Less Precise Can Be More Reliable: A Systematic Evaluation of Quantization's Impact on VLMs Beyond Accuracy

Anonymous Authors[1]

Abstract

Vision-Language Models (VLMs) such as CLIP have revolutionized zero-shot classification and safety-critical tasks, including Out-of-Distribution (OOD) detection. However, their high computational cost hinders efficient real-world deployment. While quantization is a standard solution for efficiency, its broader impact on reliability metrics beyond simple Top-1 accuracy remains critically under-explored. In this study, we conduct a large-scale evaluation of VLM quantization across a comprehensive experimental suite of over 700k evaluation runs with varying configurations. We find that, contrary to the assumption that quantization's noise degrades performance, it can simultaneously improve accuracy, calibration, OOD detection, and robustness to noise, though not to covariate shift or spurious correlations. We leverage these counterintuitive findings to characterize the mechanics of quantization beyond simple regularization: we show that quantization dampens high-rank spectral components, compelling the model to rely more heavily on robust, low-rank features. Ultimately, this spectral filtering effect drives the observed improvements in generalization and noise tolerance, establishing a pathway to deploy faster, more reliable VLMs by utilizing quantization beyond its conventional role.

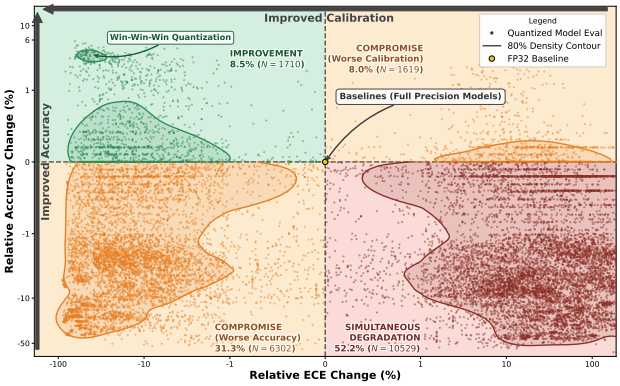

Figure 1. Quantization is not systematically destructive; A sweep of N = 20160 W8A8 quantized evaluation runs reveals distinct performance changes relative to FP32 baselines. Surprisingly, nearly 40% of evaluations show improved calibration (lower ECE), including a significant portion (8.5%) where both zero-shot accuracy and calibration improve simultaneously. Low Density around the origin suggests that Quantization is very impactful. Note that scales are logarithmic.

1. Introduction

Vision-Language Models (VLMs), particularly CLIP (Radford et al., 2021), have revolutionized computer vision through their remarkable generalization capabilities. Their powerful zero-shot performance has made them a go-to model for safety-related tasks, particularly out-of-distribution (OOD) detection. And trustworthy AI systems require deep learning models to be developed with a focus on robustness and resilience to ensure reliable and safe deployment (Hendrycks et al., 2021b; Arnez Yagualca, 2023). Consequently, a significant body of work has emerged to rigorously benchmark their reliability, with large-scale evaluations like OpenOOD and ImageNet-X specifically designed to probe the OOD robustness of foundation models (Zhang et al., 2024; Noda et al., 2025; Miyai et al., 2025a; Mayilvahanan et al., 2023; Tu et al., 2023).

However, the substantial computational requirements of these models pose a significant barrier to real-world deployment. To mitigate this, model quantization has become the standard compression method, reducing memory and computational overhead by lowering the precision of the model's weights (Courbariaux et al., 2015; Esser et al., 2020). These two domains (reliability and efficiency) have evolved in parallel, creating a critical blind spot. While the community has extensively characterized the intrinsic robustness of full-precision CLIP, it remains largely unknown if these safety-critical properties survive the aggressive, lossy compression of quantization. Our work addresses this gap.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Despite quantization's popularity, current evaluation practices suffer from a critical limitation: an overwhelming focus on top-1 accuracy while neglecting essential reliability considerations. The oversight becomes more concerning when compressed models are deployed in real-time systems, where reliability failures would have immediate consequences.

To address this gap, we move beyond conventional accuracy-focused evaluation and ask:

How does quantization impact VLMs' reliability?

To answer this question, we systematically evaluate quantization's impact across four critical reliability dimensions: (1) Robustness to quantization noise, examining stability across bit-widths and methods; (2) Uncertainty quality and calibration, evaluating whether models provide reliable confidence estimates; (3) Out-of-distribution detection, evaluating whether the model can distinguish between two to more semantic distribution of samples; (4) Distribution shift robustness, assessing performance under realistic data variations and noise; and finally (5) robustness to spurious correlation, assessing if quantized models rely more or less on spurious correlation for accurate classification.

We evaluate these five dimensions by applying a suite of 16 quantization techniques to 10 pre-trained, non-generative vision-language models[1], and testing them on rigorous benchmarks designed to assess accuracy calibration, OOD detection, and data-shift and spurious robustness. Our analysis reveals interesting findings, partially illustrated in Figure 1, showing that quantization has a measurable impact beyond accuracy, affecting all reliability attributes, indicating a complex interplay between pre-training source and compression strategy. Our findings challenge the view of quantization as a simple compression or regularization tool, and suggest it can be used to improve both model efficiency and reliability. Our major contributions are:

1. **Quantized VLM Reliability Benchmarking:** Through 8k+ quantized models and 700k+ evaluation runs, we uncover a counter-intuitive reliability compromise depending on models: quantization actively improves calibration, OOD detection, and noise robustness by regularizing models with sufficient data redundancy, but degrades data-shifted performance and exacerbates spurious correlations.
2. **Spectral Filtering Mechanism:** We identify the mechanism driving these shifts: quantization acts as a low-pass spectral filter. By injecting destructive noise into the high-rank, low-variance components, it forces models to rely on coarse, robust features, becoming less sensitive to noise and fine-grained semantic nuance.

[1]VLMs (e.g., CLIP) differ from generative, decoder-based LVLMs (e.g., LLaVA)(Miyai et al., 2025a)

2. Related Work

Quantization of Pre-trained VLMs. Quantizing large models involves a trade-off between data-free Post-Training Quantization (PTQ) (Jacob et al., 2018) and the more accurate but data-dependent Quantization-Aware Training (QAT) (Courbariaux et al., 2015). Applying QAT to foundational models requires fine-tuning on training or proxy datasets, creating a significant risk of catastrophic forgetting. While parameter-efficient methods like LoRA (Hu et al., 2021) help, the core tension between adaptation and forgetting remains an open problem.

Quantization as a Regularizer. Quantization is increasingly understood as an implicit regularizer (AskariHemmat et al., 2022) that can improve generalization by finding flatter, more robust minima in the loss landscape (Hochreiter & Schmidhuber, 1997; Tallec et al., 2023; Saqib et al., 2025). However, this effect has been studied only for weight-only quantization and exclusively through the lens of accuracy and domain generalization. Its impact on a wider range of reliability metrics, and how it interacts with a model's pre-training source (e.g., WIT (Radford et al., 2021) vs. LAION (Schuhmann et al., 2022)), remains largely unexplored.

Spectral Bias and Feature Granularity. Deep neural networks exhibit a well-documented "spectral bias," preferentially learning low-frequency, globally coherent functions before fitting high-frequency variations (Rahaman et al., 2019; Xu et al., 2024b). In the context of visual representations, this frequency hierarchy maps directly to feature granularity: low-rank spectral components typically encode coarse-grained, robust semantic structures (*e.g.*, global shapes), while the high-rank tail captures fine-grained details, textures, and noise (Yin et al., 2019; Huh et al., 2021; Wang et al., 2020).

Benchmarking VLM Reliability. The evaluation of VLM robustness and reliability has matured, with sophisticated benchmarks for OOD detection (OpenOOD) (Yang et al., 2022; Zhang et al., 2024; Wang et al., 2024) and spurious correlations (CounterAnimal) (Wang et al., 2024). While these benchmarks have been instrumental in characterizing full-precision models, their application to systematically study the impact of architectural interventions like quantization is nonexistent. Closely related to our work, Tu et al. (2023) evaluate CLIP's robustness to distribution shifts, OOD detection, and predictive uncertainty. In contrast, in this work, we go several steps further, considering the impact of VLM quantization, a key requirement for computationally efficient deployment.

3. Experimental Setup

3.1. Architectures and Quantization Protocol

Models: We evaluate 10 diverse architectures spanning standard CLIP (ViT and ConvNeXt(Liu et al., 2022)), SigLIP(Zhai et al., 2023), ALIGN(Jia et al., 2021), and CoCa(Yu et al., 2022). We fully describe our experimental setup in Appendix B. To provide a comprehensive view of VLM compression, we evaluate two distinct scopes reflecting different deployment paradigms: (1) **Visual-Only**, where the text encoder remains in FP32. This aligns with standard zero-shot classification, where prompts are static (e.g., fixed class vocabularies). In this regime, text embeddings are typically precomputed and cached offline; thus, the text encoder does not require quantization as it's never reused. (2) **Joint Visual-Text**, where both modalities are quantized to the same precision using the same method. This addresses dynamic open-vocabulary scenarios where prompts are generated at runtime, precluding caching and necessitating efficient real-time inference for both encoders.

Quantization Strategy: We utilize simulated quantization across 16 distinct strategies. This includes 8 PTQ methods: Simple MinMax (Jacob et al., 2018), SmoothQuant (Xiao et al., 2023), IGQ-ViT (Liu et al., 2024), QwT (Hubara et al., 2025), APQ-ViT (Ding et al., 2022), Rotation-based QuaRot (Ashkboos et al., 2024), Outlier Suppression (Wei et al., 2022), and Q-VLM (Wu et al., 2024). We further evaluate 8 QAT variants based on Learned Step Size Quantization (LSQ) (Esser et al., 2020) and standard QAT with the Straight-Through Estimator (Courbariaux et al., 2015). For these, we investigate two distillation regimes: *Contrastive-Only* and *Hybrid* (Contrastive + Feature MSE). Also QAT-LoRA (Xu et al., 2024a), Q-ViT (Li et al., 2022). Post-quantization, we apply Logit Tuning (Radford et al., 2021), optimizing the logit scale on the same proxy calibration set to recover probability calibration. All quantizations are performed on 1000 unique image-caption pairs of CC3M (Sharma et al., 2018), YFCC (Thomee et al., 2016), or SBU (Ordonez et al., 2011).

3.2. Evaluation Metrics

To systematically assess reliability beyond standard accuracy, we evaluate across three dimensions. Let f denote the reference FP32 model and q the quantized variant. Let $A(m, \mathcal{D})$ represent the Top-1 accuracy of model m on dataset $\mathcal{D}$.

1. OOD Detection. We evaluate the separation between in-distribution ($\mathcal{D}_{\text{ID}}$) and out-of-distribution ($\mathcal{D}_{\text{OOD}}$) confidence scores using **AUROC** and **FPR95**. We utilize 6 scoring functions: 3 traditional (MSP (Hendrycks & Gimpel, 2017), Energy (Liu et al., 2020), Entropy) and 3 VLM-specific methods (MCM (Ming et al., 2022), NegLabel (Jiang et al., 2023), and EOE (Cao et al., 2024)). Additionally, we report the **relative OOD degradation** (δ_{OOD}), calculated as the relative change in AUROC between f and q (following the form of Eq. (1)).

2. Calibration & Robustness. We measure **Expected Calibration Error (ECE)**(Guo et al., 2017) to assess confidence reliability on $\mathcal{D}_{\text{ID}}$. We also plot enhanced reliability diagrams that show the evolution of samples in the accuracy-confidence landscape to better visualize the impact of quantization. For robustness, we report the relative accuracy change $\delta(\mathcal{D})$ using Eq. (1) under natural shifts (ImageNet-A (Hendrycks et al., 2021c), ImageNet-R (Hendrycks et al., 2021a), ImageNet-V2 (Recht et al., 2019), and ImageNet-Sketch (Wang et al., 2019)) and synthetic corruptions (CIFAR-10-C (Hendrycks & Dietterich, 2019) at Severity 3). Severity values and further details on our full experimental setup are better explained in Appendix B.

$$\delta(\mathcal{D}) = \frac{A(f, \mathcal{D}) - A(q, \mathcal{D})}{A(f, \mathcal{D})} \tag{1}$$

3. Spurious Impact Metrics. To isolate the specific bias amplified by quantization from general performance loss, we formulate two derived metrics. These extend standard group disparity measures (Sagawa et al., 2020) and concepts of compression-induced forgetting (Hooker et al., 2020) to quantify relative shifts in spurious reliance. We use these metrics to shift the analytical focus from absolute performance on counter-intuitive samples to the *differential impact of quantization on spurious feature correlation*:

Delta Relative Spurious Gap (ΔRSG): Defined as $\text{RSG}(q) - \text{RSG}(f)$, where the Relative Spurious Gap for a model m is $\text{RSG}(m) = \frac{A(m,\mathcal{D}_{\text{N}}) - A(m,\mathcal{D}_{\text{C}})}{A(m,\mathcal{D}_{\text{N}})}$. Here, $\mathcal{D}_{\text{N}}$ and $\mathcal{D}_{\text{C}}$ denote the Natural and Counter-intuitive subsets, respectively. A positive value indicates that quantization disproportionately widens the gap.

Added Vulnerability (Vuln_{add}): Defined as $\delta_{\text{C}} - \delta_{\text{N}}$, where $\delta_{\mathcal{D}}$ is the relative degradation calculated via Eq. (1) for the specific subset. A positive value signifies that the quantized model has degraded more significantly on counter-intuitive examples, implying an increased reliance on spurious features.

4. Results

Having established our experimental protocol, we now systematically evaluate quantization's impact across five critical reliability dimensions. We begin by examining zero-shot accuracy to establish quantization viability in Section 4.1, and Figure 12, then proceed to analyze these reliability metrics. In Section 5, we reveal the spectral filtering mechanism that unifies these seemingly contradictory findings.

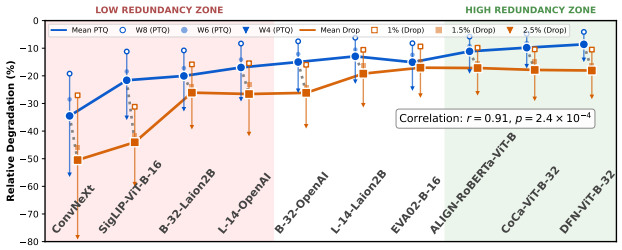

Figure 2. Available parameter redundancy indicates PTQ quantizability. A strong correlation between degradation under PTQ (blue) and weight dropout (orange) indicates that redundancy is the primary robustness buffer. Models in the High Redundancy Zone (right) utilize excess representational capacity to absorb quantization noise, whereas Low Redundancy models (left) lack this margin, leading to a significant loss of information.

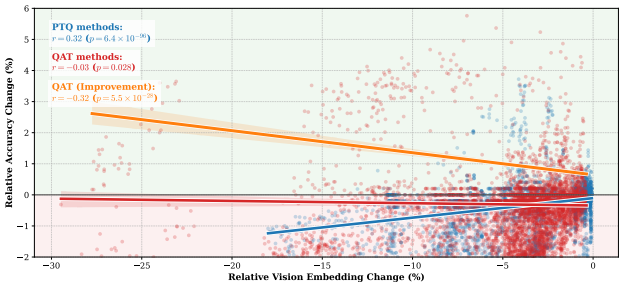

Figure 3. Vision embedding similarity between the original and quantized model is a good indicator of information retention for PTQ. However, QAT requires altering embeddings to values that are more quantizable and thus diverge from the baseline embedding.

4.1. The Zero-Shot Accuracy Landscape under Quantization

Less precise can be more accurate. The results (Figure 1) suggest that weight and activation quantization-aware training (QAT) can aid generalization by driving models toward flatter minima. However, we find that the success of this approach depends heavily on the interplay between data quality and model size. While QAT generally smoothes the optimization landscape compared to the brittle solutions found by Post-Training Quantization (PTQ), a divergence appears in the capacity-constrained `Base-32` architecture (please refer to the Appendix Figure 12 for the per-model quantizibility). Here, the model trained on high-quality WIT data quantizes well, whereas the variant trained on noisy LAION data suffers significant degradation. This suggests that the smaller `Base-32` lacks the redundancy to withstand the combination of LAION's inherent label noise and injected quantization noise. In contrast, `Large-14` models don't exhibit this difference, remaining robust and well-quantized across both WIT and LAION datasets, suggesting the model capacity is better for quantizability. Ultimately, reduced precision can improve accuracy depending on how higher and flatter the new minima are. We develop this idea in Appendix G, and plot per-quantization-method results in Appendix Figure 14.

Data curation and scale generate the redundancy that quantization taps into. We identify a strong linear correlation ($r = 0.91$) between static quantization degradation and weight dropout sensitivity, confirming that resilience is a function of available parameter redundancy (Figure 2). Crucially, this redundancy is not solely determined by model size, but is heavily modulated by pre-training data quality. Models trained on massive but uncurated, noisy datasets such as `SigLIP` and `CLIP-ViT-B-32-Laion2B` fall into the *Low Redundancy Zone* (left), suggesting that their capacity is fully exhausted fitting the high-entropy noise of the data, leaving no "safety margin" for quantization. In contrast, models trained on high-quality, filtered data like `DFN-ViT-B-32` inhabit the *High Redundancy Zone* (right), outperforming even larger architectures (e.g., `L-14-Laion2B`). This indicates that data filtering in the DFN model produces cleaner, more redundant parameters, allowing smaller models to survive quantization better than larger models trained on noisy scrapes.

Fidelity to the baseline is a valid proxy for PTQ success, but a misleading constraint for QAT. As shown in Figure 3, we measured the visual fidelity by comparing the cosine similarities of the baseline and quantized visual encoders. We observe a statistically significant positive correlation ($r = 0.32$) between vision embedding similarity and accuracy for PTQ methods. This confirms that for post-training quantization, preserving the original manifold is critical; any deviation is effectively "noise" that degrades performance. In sharp contrast, QAT decouples performance from fidelity ($r = -0.03$), demonstrating that fine-tuning allows the model to find new, quantization-friendly minima that are distinct from the FP32 basin. Most strikingly, the subset of models that *improve* over the baseline (Orange Regression, $r = -0.32$) exhibit a significant *negative* correlation; meaning that the highest accuracy gains are achieved not by mimicking the original model, but by actively diverging from its embedding space. This suggests that QAT improves generalizability (Saqib et al., 2025) by shedding brittle features. We further address the PTQ vs QAT performance in Appendix E and the counter-intuitive impact of task complexity on quantization in Appendix D.

4.2. Calibration and Uncertainty Quality

Pre-training data quality is correlated to better calibration under quantization. The impact of quantization on predictive uncertainty (Figure 4) reveals that precision loss does not systematically degrade model trust; Its impact is heavily modulated by the pre-training source. Models pre-

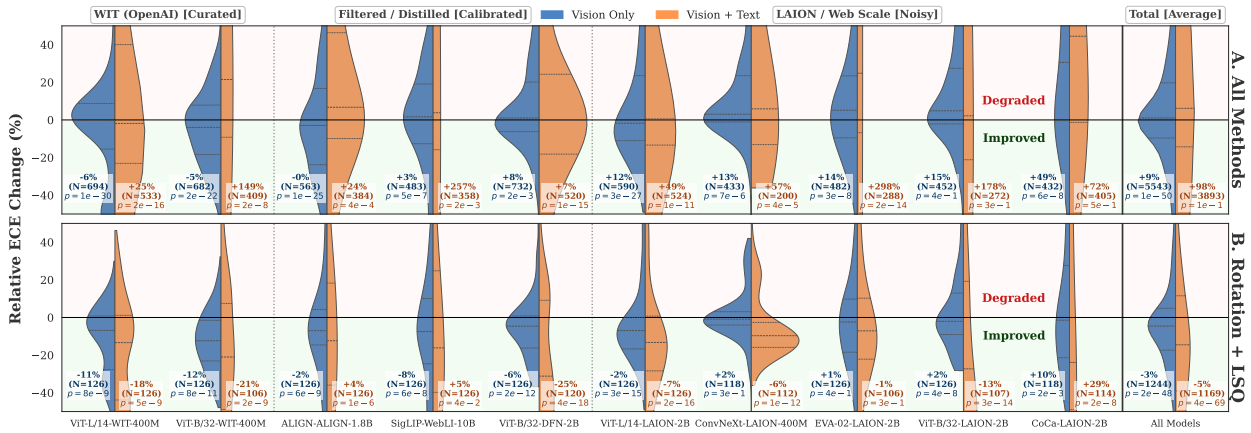

Figure 4. Direct impact of successful W8A8 quantizations on calibration. Models are ordered by relative ECE Change. Results indicate a correlation between pre-training data quality and post-quantization calibration degradation: curated sources (WIT) benefit from quantization-induced regularization, while noisy web-data (LAION) leads to increased calibration error. The "Total" distribution demonstrates that vision-only quantization is generally more robust to quantization-induced calibration degradation, whereas text quantization is significantly more brittle. Crucially, comparing Panel A to Panel B confirms that quantization training is necessary to maintain or improve calibration, as PTQ quantization generally renders the model's logit scales obsolete, a problem that is even more pronounced when the text encoder is also quantized.

trained on curated data (*e.g.*, WIT, DFN) frequently exhibit *improved* calibration, as quantization noise provides stochastic regularization that dampens the overconfidence typical of foundation models. Conversely, models trained on noisy web data (LAION) suffer significant degradation (+49% ECE), as precision loss pushes their already diffuse representations into further misalignment. Crucially, we identify text quantization as a primary point of failure, driving a catastrophic +98% increase in calibration error (Panel A). This occurs because zero-shot classification relies on the dot product of two independently quantized manifolds, rendering the pre-trained logit scales obsolete. However, a QAT methods (Panel B) help mitigate this with backpropagation. Effectively absorbing scale distortions and achieving a global -5% reduction in ECE compared to the FP32 baseline, even more so when the text encoder is also quantized.

Logit Scale Tuning: This is a form of temperature scaling (Guo et al., 2017) that recalibrates the pre-trained logit scale, rendered obsolete by quantization. To visualize this, we replace static reliability bars with bin trajectories (Figure 5), tracking the evolution of confidence buckets from the baseline. This reveals distinct pathologies: curated models (WIT) start systematically under-confident, while noisy models (LAION) are over-confident. Optimizing logit scales on proxy data while freezing the backbone corrects this geometric mismatch. As shown by the dotted trajectories, this simple scalar realignment effectively pulls predictions toward the diagonal. Consequently, fully quantized models achieve an average ECE of 1.1%, significantly outperforming their own FP32 baselines (6.9%) that have their logit scales trained on the whole pre-trained dataset.

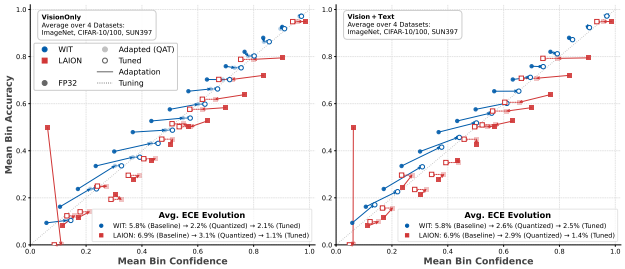

Figure 5. **Calibration Trajectories.** We modify standard reliability diagrams to plot the *trajectory* of confidence bins from FP32 (solid) → QAT (faded) → Logit Tuning (hollow). This visualization reveals how quantization and tuning dynamically correct the systematic under-confidence of curated models (WIT, blue) and over-confidence of noisy models (LAION, red), allowing quantized models to surpass FP32 calibration.

However, the success of this tuning depends sensibly on the quality of the proxy data. As shown in Figure 15 (Left), while accuracy degradation (around 6.5%) remains consistent across proxy sources, calibration is far more sensitive to these changes. The curated CC3M dataset yields a statistically significant improvement in ECE retention ($p < 0.001$) compared to noisy alternatives like YFCC. Furthermore, Figure 15 (Right) reveals a strong "repairability" correlation: models with the most severe post-quantization misalignment (high ECE after being quantized) experience the largest gains from tuning. This further supports the idea of post-quantization logit obsolescence: as the logit scale is trained for millions of iterations and quantization heavily impacts not only the internal representations but also the model's final accuracy, the logit scale becomes misaligned with the altered distribution of quantized outputs.

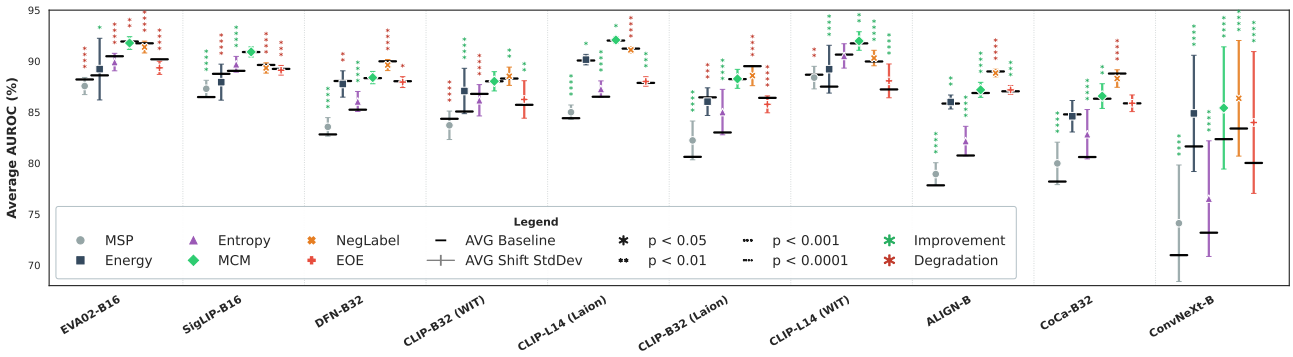

Figure 6. Impact of successful W8A8 quantization on OOD Detection (AUROC). Average AUROC across quantization methods (higher is better). QAT methods (center, right) maintain OOD performance for the LAION model, despite this model suffering from significant accuracy and calibration degradation. VLM-specific OOD methods consistently outperform classic methods. Vertical lines represent the maximum improvement and degradation relative to the full precision baseline of that experiment. Refer to Figure 20 for $FPR@95$.

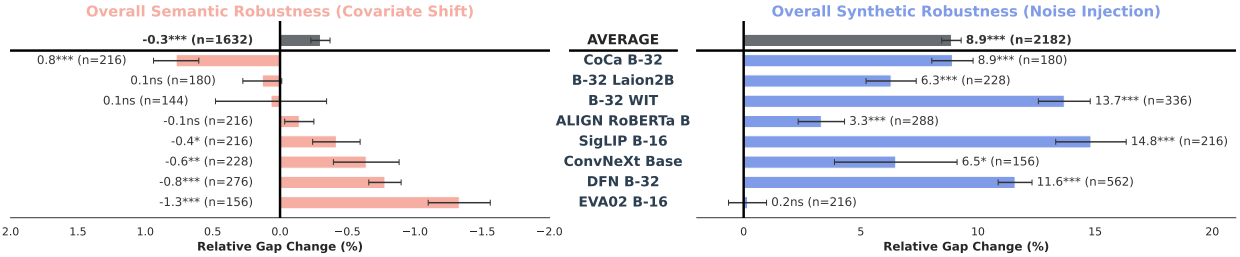

Figure 7. Quantization Impacts On Robustness; A comparison of relative robustness gap changes under successful W8A8 quantization (less than 2% degradation on the non-shifted dataset) across varied VLM architectures and quantization methods reveals that while quantization generally slightly degrades semantic robustness (covariate shift, Left), it largely improves robustness against synthetic noise, (Right). Annotations indicate mean change, error, and significance $p < 0.05, 0.01, 0.001$, or non-significance (ns).

4.3. Out-of-Distribution Detection: Resilience and Regularization

OOD In The VLM Era. We observe a consistent hierarchy where VLM-specific methods leveraging textual anchors (MCM, NegLabel) outperform unimodal baselines (Wang et al., 2023). However, on standard benchmarks, the gap to energy-based scores is narrow, suggesting raw logit norms capture sufficient semantic distance in foundation models. Notably, for architectures such as `SigLIP` and `DFN-B32`, specific quantization configurations yield small but statistically significant AUROC improvements Figure 6, suggesting that discretization may filter high-frequency noise and clarify decision boundaries. However, since accuracy is also strongly correlated with OOD detection, degradations in the former also impact the latter.

Re-coupling Accuracy and OOD Detection. Statistical analysis ($N = 466k$) indicates a strong coupling between classification accuracy and OOD detection. As shown in Appendix Figure 16 (A), metrics are perfectly correlated ($r = 0.994$) (Wang et al., 2024), this implies that quantization impact on top-1 accuracy and on calibration will also be translated to OOD performance.

Architectural Divergence: We observe that `ConvNeXt` benefits significantly more from quantization than Transformer-based architectures. We hypothesize that quantization acts as a spectral filter for CNNs, dampening the high-frequency, textural features that typically lead to OOD overconfidence in full precision. By stripping away this spurious precision, quantization effectively acts as a denoiser, enhancing the separability of the semantic manifold (see Appendix F.1 for detailed separability analysis).

4.4. Robustness to Covariate Shift

The Robustness Dichotomy: Noise vs. Semantics. As illustrated in Figure 7, quantization introduces a fundamental divergence in how VLMs respond to distribution shifts. We observe a stark contrast between synthetic robustness (resilience to noise, blur, and weather) and semantic robustness (resilience to style changes and natural shifts like ImageNet-A/R). This dichotomy suggests that quantization is not merely a compression operation, but a structural filter that fundamentally alters the model's feature dependence.

Synthetic Robustness: Quantization as a Low-Pass Filter. The most significant finding is the universal and massive

Table 1. Impact of quantization on spurious correlations (%). We report ΔRSG 3.2 (increase in Relative Spurious Gap) and Vuln. 3.2 (Added Vulnerability). Rot+LSQ does not significantly impact spurious correlations at 8-bit and 6-bit settings. Realized on the Counter-Animal Dataset (Wang et al., 2024).

		Bit-width (W/A)		
Method	**Metric**	**8/8**	**6/8**	**4/8**
Simple PTQ	ΔRSG	$2.6\pm0.2^{***}$	$2.8\pm0.3^{***}$	$12.5\pm0.6^{***}$
	Vuln.	$3.0\pm0.3^{***}$	$3.2\pm0.3^{***}$	$10.3\pm0.4^{***}$
QAT (Contr.)	ΔRSG	$1.6\pm0.2^{***}$	$2.1\pm0.3^{***}$	$9.1\pm0.5^{***}$
	Vuln.	$1.9\pm0.3^{***}$	$2.4\pm0.3^{***}$	$9.1\pm0.4^{***}$
Rot+LSQ (Ours)	ΔRSG	$\mathbf{-0.1\pm0.1}^{ns}$	$\mathbf{0.2\pm0.1}^{ns}$	$\mathbf{4.0\pm0.3}^{***}$
	Vuln.	$\mathbf{-0.2\pm0.1}^{ns}$	$\mathbf{0.2\pm0.1}^{ns}$	$\mathbf{4.4\pm0.3}^{***}$

improvement in synthetic robustness (Right Panel), with an average relative gain of $+8.9\%$ across all architectures. Synthetic corruptions fall into two categories: those that introduce high-frequency noise (e.g., Gaussian noise) and those that remove it (e.g., defocus blur). By discretizing the network weights, quantization effectively limits the model's capacity to represent fine-grained, high-frequency features. This forces the model to rely on coarse, low-frequency spectral components (such as global shape and structure). Consequently, the quantized model becomes robust to high-frequency perturbations, whether they are additive noise (which is ignored) or blur (which removes details that the model relies less on).

Semantic robustness suffers from feature erasure. Conversely, this same filtering mechanism proves detrimental under semantic covariate shift (Left Panel). When facing natural distribution shifts involving style transfer or hard examples (ImageNet-A/R/Sketch), we observe a consistent degradation relative to the FP32 baseline. Semantic robustness often relies on fine-grained, high-frequency features to distinguish objects in data-shift contexts; quantization compresses these nuanced signals into shared bins, effectively erasing the discriminatory detail required for these semantic edge cases. Notably, `EVA02 B-16` suffers the most severe degradation, suggesting that architectures optimized for high-density information packing are more susceptible than generative-based architectures like `CoCa`.

4.5. Spurious Correlations

The Bias Amplification Risk. While quantization improves robustness to synthetic noise, Table 1 reveals a concerning side effect regarding another aspect of reliability. Quantization methods such as PTQ and QAT (families) significantly exacerbate reliance on spurious correlations, increasing the Relative Spurious Gap (RSG) by $+2.7\%$ at W8A8 and a massive $+12.5\%$ at W4A8 ($p < 0.001$). This suggests that the filtering mechanism described in the previous section cuts both ways: by dampening fine-grained signals, naive inadvertently suppresses the subtle invariant features required to distinguish core concepts from their contexts (e.g., distinguishing a "wolf" from a "snowy background"). Consequently, the model reverts to "heuristic" reasoning, latching onto high-magnitude background correlations that survive the discretization process.

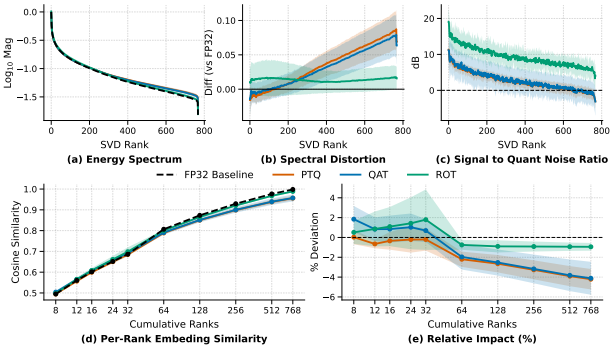

Figure 8. Quantization Impact on the Feature Space. (a–c) Spectral analysis reveals that standard quantization disproportionately degrades low-energy components, causing significant spectral distortion and collapsing the SQNR at higher ranks. (d–e) Show how quantization forces the model to do more with less, achieving higher similarity at lower ranks, but worse with more. Conversely, rotation-based methods ROT mitigate this increasingly important quantization noise due to their smaller step size, preserving good semantic alignment with the baseline, where standard PTQ and QAT drift significantly at higher ranks.

Rotation Mitigates Bias. This bias amplification is not inevitable. *Rotation + LSQ* method is the most robust quantization technique to spurious correlation exacerbation, yielding statistically insignificant changes to spurious vulnerability at W8A8 and W6A8 (-0.2 n.s.). By rotating the activation space prior to quantization, this method aligns the semantic manifold with the quantization grid, preventing the "outlier collapse" that typically destroys feature subtleties. However, at extreme compression rates (W4A8), even robust methods succumb to bias (+3.7%), indicating a hard limit where the bit-depth to suppress quantization noise.

5. Mechanics of Quantization: Spectral Filtering

To understand the mechanisms driving the observed trade-offs between reliability metrics, we analyze the spectral properties of the quantized representations. We hypothesize that although quantization injects uniform noise in the value domain, it operates as a non-uniform spectral filter in the feature domain, disproportionately impacting components based on their variance. We performed Singular Value Decomposition (SVD) on the penultimate visual layer of each model in the ImageNet validation set across 3 independent

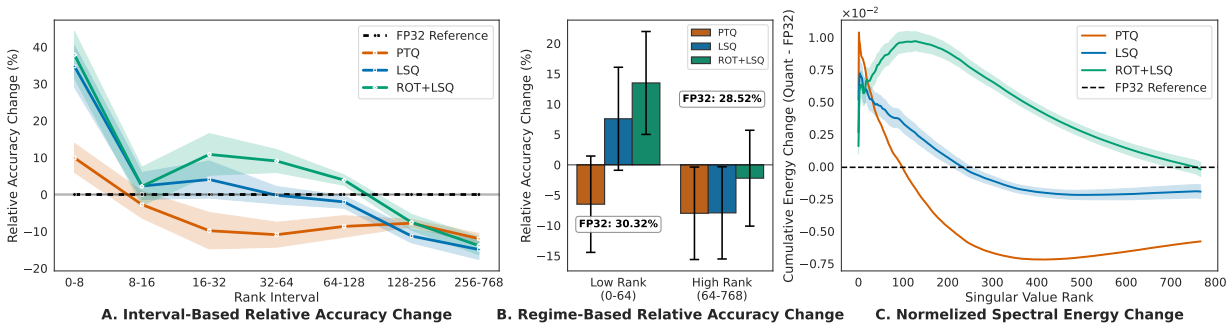

Figure 9. Impact of quantization methods on subspace accuracy and spectral energy. (A) Relative accuracy change across rank intervals compared to the FP32 baseline. (B) Direct comparison between low-rank (0–64) versus high-rank (64–768) regimes' accuracy change. (C) Deviation in cumulative spectral energy relative to the FP32 reference, showing energy shift across singular value ranks from higher ranks to lower ranks. Results stem from using 10 full independent runs on `CLIP-ViT-B-32-WIT` on ImageNet1k.

runs, aggregating results across all models and grouping them by method family (PTQ, QAT, or ROT). Our analysis proceeds in two stages: first, projecting quantized features onto the original FP32 basis to quantify signal preservation (Figure 8), and second, recomputing the SVD on the quantized features to observe information reorganization (Figure 9). Please refer to the Appendix C for a detailed methodology description.

Passive Spectral Filtering. Figure 8 demonstrates that the Signal-to-Quantization-Noise Ratio (SQNR) decays monotonically with increasing rank. Because the quantization step size is fixed, spectral components with magnitudes comparable to the discretization grid (typically high-rank, low-variance features) are dominated by rounding errors (Figure 8 b-c). This effectively functions as a low-pass filter, preserving the dominant semantic structure while attenuating fine-grained details (Figure 8 d-e). This mechanism elucidates the synthetic robustness improvements reported in Section 4.4: by truncating the spectral tail where both additive noise and fine-grained degradation manifest, quantization desensitizes the model to these corruptions as it adapts (with QAT) to rely on coarser, generalist features.

Active Subspace Concentration. Examining the intrinsic discriminability of the quantized latent space (Figure 9) reveals an active compensation mechanism. In the dominant principal components (Ranks 0–8), quantized models surprisingly achieve higher subspace accuracy than their FP32 baselines (Figure 9A). This suggests a "coarse-grained compensation" effect with QAT: limited representation capacity compels the model to concentrate discriminative information in the most robust, high-variance dimensions, which are less affected by rounding noise. This concentration directly correlates with improved OOD detection, as the suppression of ambiguous, low-variance signals clarifies in-distribution decision boundaries by reducing the noise floor of the representation (see significant gains on ConvNex F.1).

The Semantic Trade-off and Mitigation. However, this spectral filtering imposes a cost. Figure 9B highlights a significant degradation in accuracy for high-rank subspaces (Rank 64+). Since these dimensions encode the subtle variations necessary for resolving fine-grained classifications and adapting to covariate shifts (e.g., style changes), their suppression explains the reduced performance on semantic robustness benchmarks and the increased susceptibility to spurious correlations. Notably, rotation-based methods mitigate this trade-off. As evidenced by the SQNR retention in Figure 8, rotation aligns the activation distribution with the quantization grid. This prevents premature signal decay in the intermediate spectral region (Ranks 8–64), allowing these methods to retain critical semantic detail without sacrificing the regularization benefits of low-rank concentration.

6. Conclusion

This work challenges the prevailing belief that quantization is merely a compromise between efficiency and accuracy. Through the largest systematic evaluation of quantized VLMs to date, we demonstrate that discretization functions as an unintended but powerful spectral regularizer. By suppressing high-rank, low-variance components that are most sensitive to rounding noise, quantization compels the model to rely more on coarse-grained features. When a task requires generalization over specificity, this lower precision can simultaneously improve calibration, uncertainty estimation, and robustness to synthetic noise. However, this same mechanism could erase the fine-grained nuances required for semantic robustness and amplify spurious correlations. Crucially, we identify that the margin between beneficial regularization and destructive collapse is governed by the information quality of pre-training data and the model's capacity, which provides the necessary redundancy to buffer the spectral bias exacerbated by quantization. Overall, this work provides guidance for practitioners and a paramount research direction for utilizing quantization beyond efficiency.

Impact Statement

This work provides a systematic evaluation of how model quantization affects the reliability of Vision-Language Models (VLMs) and systems that use them. On the positive side, our discovery that quantization can improve OOD detection and noise robustness offers a pathway to deploying safer AI systems in resource-constrained environments. However, we also reveal a potential ethical risk: the spectral filtering effect can increase vulnerability to spurious correlations and thus amplify unfair and unethical stereotypes that may already be present in the internet-scrapped training sets (Birhane et al., 2021), as it have been previously shown that compression techniques could negatively impact fairness (Hooker et al., 2021). By identifying rotation-based quantization as more robust to this degradation, we provide practitioners with general guidelines for mitigating it. Overall, this paper presents work aimed at advancing the field of Machine Learning, specifically by improving VLM's frugality, which we consider to be a net positive. There could still be more potentially negative societal consequences of our work, none of which we foresee at the moment.

References

Arnez Yagualca, F. A. *Deep neural network uncertainty runtime monitoring for robust and safe AI-based automated navigation*. Theses, Université Paris-Saclay, December 2023. URL https://theses.hal.science/tel-04672736.

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. Quarot: Outlier-free 4-bit inference in rotated llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

AskariHemmat, M., Hemmat, R. A., Hoffman, A., Lazarevich, I., Saboori, E., Mastropietro, O., Sah, S., Savaria, Y., and David, J.-P. Qreg: On regularization effects of quantization. *arXiv preprint arXiv:2206.12372*, 2022.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. doi: 10.48550/arXiv.2110.01963. URL https://arxiv.org/abs/2110.01963.

Cao, C., Zhong, Z., Zhou, Z., Liu, Y., Liu, T., and Han, B. Envisioning outlier exposure by large language models for out-of-distribution detection. In *International Conference on Machine Learning (ICML)*, 2024.

Courbariaux, M., Bengio, Y., and David, J.-P. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems (NIPS)*, volume 28, 2015.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Ding, Y. et al. Towards accurate post-training quantization for vision transformer. In *ACM*, 2022.

Esser, S. K., McKinstry, J. L., Bablani, D., Mallya, A., Appuswamy, R., and Rath, D. Learned step size quantization. In *International Conference on Learning Representations (ICLR)*, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning (ICML)*, pp. 1321–1330. PMLR, 2017.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Hvilshoj, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021a.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021b.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021c.

Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., and Bethge, M. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Hooker, S., Courville, A., Clark, G., Dauphin, Y., and Frome, A. What do compressed deep neural networks forget? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2093–2102, 2020. URL https://arxiv.org/abs/1911.05248.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models, 2021. URL https://arxiv.org/abs/2010.03058.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. Published at the International Conference on Learning Representations (ICLR) 2022.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8710–8719, 2021.

Hubara, I. et al. Quantization without tears: Improving post-training quantization for large-scale models. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Huh, M., Mobahi, H., Zhang, R., Cheung, B., Agrawal, P., and Isola, P. The low-rank simplicity bias in deep networks. In *TMLR*, 2021.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2704–2713, 2018.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Jiang, J. et al. Detecting out-of-distribution data through the lens of adversarial examples. In *CVPR (Note: Verify specific NegLabel paper source if different, e.g., "Negative Prompting")*, 2023.

Kar, P., Arık, S. O., Choi, D., Bhattacharjee, B., Lien, A.-T., and Pfister, T. Locoop: Few-shot out-of-distribution detection via prompt learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems (NIPS)*, volume 31, 2018.

Li, Y. et al. Q-vit: Accurate and efficient low-bitwidth vision transformer. In *Advances in Neural Information Processing Systems*, 2022.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21464–21475, 2020.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Liu, Z. et al. Post-training quantization for vision transformer via instance-aware group quantization. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

Mayilvahanan, P., Wiedemer, T., Rusak, E., Bethge, M., and Brendel, W. Does CLIP's Generalization Performance Mainly Stem from High Train-Test Similarity? *arXiv preprint arXiv:2310.09562*, 2023.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Miyai, A., Yang, J., Zhang, J., Ming, Y., Lin, Y., Yu, Q., Irie, G., Joty, S., Li, Y., Li, H., Liu, Z., Yamasaki, T., and Aizawa, K. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *Transactions on Machine Learning Research*, 2025a. URL https://arxiv.org/abs/2407.21794. Survey paper.

Miyai, A., Yang, J., Zhang, J., Ming, Y., Lin, Y., Yu, Q., Irie, G., Joty, S., Li, Y., Li, H., et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. In *TMLR*, 2025b.

Noda, S., Miyai, A., Yu, Q., Irie, G., and Aizawa, K. A benchmark and evaluation for real-world out-of-distribution detection using vision-language models. *arXiv preprint arXiv:2501.18463v1*, 2025.

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, pp. 5301–5310, 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.08731.

Saqib, J., Hieu, L., and Mathieu, S. QT-DoG: Quantization-aware training for domain generalization. In *International Conference on Learning Representations (ICLR)*, 2025.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Shao, W., Zhao, L., He, Z., Jiao, Z., Chen, P., and Ng, K.-T. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=KQP1Vfb2g3.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Slyman, E., Kanneganti, A., Hong, S., and Lee, S. You never know: Quantization induces inconsistent biases in vision-language foundation models. *arXiv preprint arXiv:2410.20265*, 2024. Workshop paper at NeurIPS 2024 RBFM.

Tallec, C., Blier, L., and Ollivier, Y. Revisiting the regularization effect of quantization. *arXiv preprint arXiv:2310.03113*, 2023.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Bency, D., and Li, X. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

Tu, W., Deng, W., and Gedeon, T. A closer look at the robustness of contrastive language-image pre-training (clip). *Advances in Neural Information Processing Systems*, 36: 13678–13691, 2023.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Wang, H., Wu, X., Huang, Z., and Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8684–8694, 2020.

Wang, H., Li, Y., Yao, H., and Lin, X. Clipn: Zero-shot ood detection via text-image alignment. In *ICCV*, 2023.

Wang, Q., Lin, Y., Chen, Y., Schmidt, L., Han, B., and Zhang, T. A sober look at the robustness of CLIPs to spurious features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Wei, X. et al. Outlier suppression: Pushing the limit of low-bit transformer quantization. In *Advances in Neural Information Processing Systems*, 2022.

Wu, H. et al. Q-vlm: Post-training quantization for vision-language models. In *Advances in Neural Information Processing Systems*, 2024.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, pp. 38087–38101. PMLR, 2023.

Xu, Y. et al. Qa-lora: Quantization-aware low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2024a.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2024b.

Yang, J., Zhou, K., and Liu, Z. Full-spectrum out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16293–16302, 2022.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 2024.

Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness

in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., and Li, H. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024.

Appendix

A. Limitations

Our study, while comprehensive, has certain limitations.

- **Simulated Quantization:** Unless explicitly stated otherwise, all experiments utilize simulated (or "fake") quantization. We acknowledge that performance in actual deployment may vary depending on the specific hardware architecture and quantization frameworks employed. We further detail why simulated quantization is imperative in our study in Section F.3, and we verify the validity of all quantization using safeguards detailed in Section F.3.

- **Proxy Dataset:** All quantization methods were adapted to caption-based proxy datasets. We did not explore the effect of quantization of fine-tuned VLMs on downstream tasks. Nor do we include VFMs, LLVMs, or other architectures such as LLaVa and Qwen that build on top of VLMs like CLIP, as that would have uncontrollably expanded the scope of this work.

- **Dynamic Quantization:** In our work, we quantize all linear layers without prior analysis. We did not exclude some linear layers from quantization to gain performance.

- **Pure Zero-shot:** Some important baselines are missing from this work, some due to the complexity of implementation, and others due to relying on ID data, e.g., Mahalanobis (Lee et al., 2018), LoCoOp (Kar et al., 2023), OmniQuant (Shao et al., 2023).

- **Robustness to Adversarial Attacks:** Although typically categorized under model robustness, we focus exclusively on safety concerns rather than security. Consequently, our scope excludes malicious manipulations and addresses only unintentional, undesirable model behaviors.

B. Extended Experimental Setup

This appendix provides the granular configuration details for our large-scale study, which encompasses over 700,000 unique evaluation data points. *Note that we quantized all linear layers, even when they are encapsulated (e.g., in multi-head attention), as well as all convolution layers. The only exceptions to quantization are the absolute last projection layers of each encoder.*

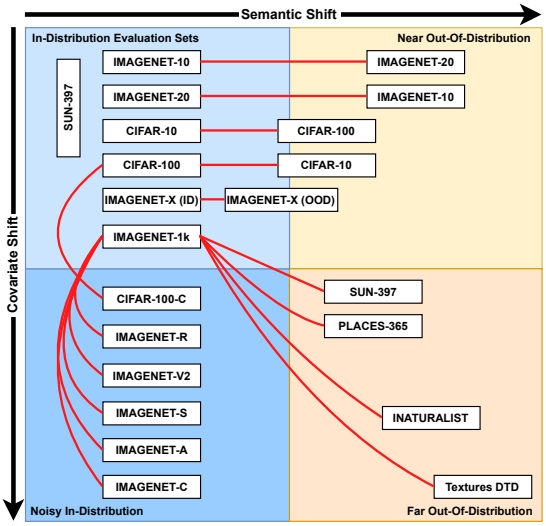

Figure 10. Systematic benchmark suite categorized by Covariate and Semantic Shift. Arrows indicate the ID reference dataset for OOD detection.

B.1. Benchmark Dataset Taxonomy

Our benchmark suite systematically covers different forms of distribution shift, and we use the same datasets that are recognised and used by the community (Yang et al., 2024; 2023; Miyai et al., 2025b). For more details on dataset preparation for OOD detection, please refer to citep . They are categorized along two dimensions: *Covariate Shift* and *Semantic Shift*, as shown in Figure 10. We also use a specialized dataset for spurious correlations (Hochlehnert et al., 2025).

B.2. Proxy Datasets and Training

For methods that require data (QAT, LSQ, and Calibration), we use three primary proxy datasets: **CC3M** (Conceptual Captions), **YFCC**, and **SBU** Captions.

- **Sample Volume:** For each proxy scenario, we stream **1,000 image-caption pairs** from WebDataset shards to serve as the training/calibration distribution. We explain our motivation here F.2

- **Hyperparameters:** Optimization is conducted using AdamW with a base learning rate of 10^{-6} for model weights. For LSQ, quantization step-sizes (Δ) are optimized with a higher learning rate of 10^{-4} to ensure rapid convergence of the clipping thresholds.

B.3. Standard and OOD Benchmarks

Full Evaluation: All downstream evaluations are performed on the **full validation sets** of each dataset (e.g., all 50,000 images of ImageNet-1k, except for the MOS-subsets to follow thier benchmark).

- **Semantic Shift (MOS):** For Far-OOD detection, we utilize the **Maximum Over Softmax (MOS)** benchmark protocol (Huang & Li, 2021). We evaluate ImageNet-1k as the In-Distribution (ID) source against three distinct OOD targets: **MOS-SUN397**, **MOS-Places365**, and **MOS-iNaturalist**, alongside the Describable Textures Dataset (**DTD**).
- **ID-OOD Pairs:** We evaluate 9 specific ID-OOD pairings, including semantic swaps (CIFAR-10 vs. 100) and curated subsets (ImageNet-10 vs. 20 and ImageNet-500 ID vs. OOD).

B.4. Robustness and Corruption Details

We evaluate robustness using two distinct pipelines:

- **Natural Distribution Shifts:** We utilize the full test suites of ImageNet-V2, ImageNet-A (Adversarial), ImageNet-R (Rendition), and ImageNet-Sketch.
- **Synthetic Corruptions (CIFAR-10-C):** We implement a pipeline using the `imagecorruptions` library, specifically targeting **Severity Level 3** (on a scale of 1–5) to represent significant but non-destructive noise. The corruptions include:
  1. **Gaussian Noise:** Additive electronic noise ($\sigma = 0.08$).
  2. **Defocus Blur:** Simulating camera misfocus via a disk kernel.
  3. **Brightness:** Multiplicative intensity shift (factor = 1.5).
  4. **Contrast:** Histogram stretching/compression (factor = 1.5).

B.5. Justifying the 700k+ Evaluation Runs

The combinatorial complexity is calculated as follows:

1. **Model States:** 10 Models $\times$ 3 Precisions ($W8A8, W6A8, W4A8$) $\times$ 2 Scopes (Visual vs. Joint) $\times$ 3 Seeds $\times$ 3 Data Scenarios $\times$ 16 Method Variants $\approx$ **8,640 unique model checkpoints**.
2. **Evaluations per Checkpoint:**
   - **Zero-Shot Accuracy:** 7 datasets x 2 (Logit Tuning) = 14 evals per checkpoint.
   - **OOD Metrics:** 9 Dataset Pairs $\times$ 6 Scoring Functions (MSP, Energy, Entropy, MCM, NegLabel, EOE) = 54 evals per checkpoint.
   - **Robustness:** 4 Natural Shifts + 4 Synthetic Corruptions + 2 Baselines - 2 (ViT-L-14) = 8 evals per checkpoint.
   - **Spurious:** 2 Spurious splits = 2 evals per checkpoint.
   - **Spectral:** 1 dataset (ImageNet) x 10 intervals of ranks x 0.3 (vision-only quantization, 2 seeds instead of 3) = 4 evals per checkpoint.

Total Calculation: $8,640$ states $\times$ 82 evaluations $\approx$ $\mathbf{708,480}$ unique evaluation runs. We will upload all these results in CSV format along with our code. All benchmarks were performed on an Nvidia H200 cluster, and required $\approx 12,000$ GPU-hours.

C. Spectral Analysis Methodology

To characterize the mechanics of quantization beyond scalar metrics, we developed a two-stage spectral analysis framework. This section details the mathematical formulation for the plots presented in the main text. All analyses were conducted on the penultimate visual features of the ImageNet-1k validation set.

C.1. Passive Filtering Analysis (Fixed Basis)

This analysis measures how well the quantized model preserves the original information structure. We define $\mathbf{X} \in \mathbb{R}^{N \times D}$ as the activations of the baseline FP32 model and $\mathbf{X}_Q \in \mathbb{R}^{N \times D}$ as the activations of the quantized model.

First, we compute the singular value decomposition (SVD) of the centered baseline features:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{D \times D}$ represents the frozen orthogonal basis of the original feature space, and $\mathbf{v}_k$ denotes the k-th eigenvector. For Fixed Basis, the SVD ranks are not re-ordered as the energy change when the model is quantized.

Energy Spectrum & Spectral Distortion. We project the quantized activations onto the *original* FP32 basis $\mathbf{V}$ to measure the magnitude of the signal retained along each original principal component k. The projected magnitude $\sigma_Q^{(k)}$ is given by:

$$\sigma_Q^{(k)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((\mathbf{x}_Q^{(i)} - \boldsymbol{\mu}) \cdot \mathbf{v}_k\right)^2} \tag{3}$$

The **Spectral Distortion** (Figure 8b) is defined as the logarithmic difference relative to the baseline singular values:

$$\Delta\sigma^{(k)} = \log_{10}(\sigma_Q^{(k)}) - \log_{10}(\sigma_{FP32}^{(k)}) \tag{4}$$

A negative distortion indicates signal attenuation (damping), while positive indicates signal amplification (noise injection).

Signal-to-Quantization-Noise Ratio (SQNR). The SQNR (Figure 8c) quantifies the fidelity of the representation for each specific spectral rank. We calculate the Signal Power (P_S) and Quantization Noise Power (P_N) projected along each eigenvector $\mathbf{v}_k$:

$$P_S^{(k)} = \mathbb{E}\left[((\mathbf{X} - \boldsymbol{\mu})\mathbf{v}_k)^2\right] \tag{5}$$

$$P_N^{(k)} = \mathbb{E}\left[((\mathbf{X} - \mathbf{X}_Q)\mathbf{v}_k)^2\right] \tag{6}$$

The SQNR is reported in decibels (dB):

$$\text{SQNR}_k = 10 \cdot \log_{10}\left(\frac{P_S^{(k)}}{P_N^{(k)} + \epsilon}\right) \tag{7}$$

A collapse in SQNR_k indicates that the variance of the quantization error exceeds the natural variance of the feature, effectively destroying information at that rank.

Subspace Similarity Saturation. To measure global semantic alignment (Figure 8d–e), we reconstruct the embeddings using only the first k principal components and compute their cosine similarity to the full FP32 embedding. Let $\hat{\mathbf{x}}_k$ be the reconstruction at rank k:

$$\hat{\mathbf{x}}_k = \boldsymbol{\mu} + \sum_{j=1}^{k} ((\mathbf{X}_Q - \boldsymbol{\mu}) \cdot \mathbf{v}_j)\, \mathbf{v}_j^T \tag{8}$$

We report the average cosine similarity: $\text{Sim}_k = \text{CosSim}(\hat{\mathbf{x}}_k, \mathbf{X}_{FP32})$.

C.2. Active Subspace Analysis (Adaptive Basis)

This analysis investigates how the quantized model *reorganizes* its own latent space. Unlike the passive analysis, we compute a **new** SVD basis $\mathbf{V}_Q$ specific to the quantized model's features:

$$\mathbf{X}_Q - \boldsymbol{\mu}_Q = \mathbf{U}_Q \boldsymbol{\Sigma}_Q \mathbf{V}_Q^T \tag{9}$$

Subspace Classification Accuracy. We isolate specific spectral bands (e.g., Rank i to j) to determine which components drive the model's performance (Figure 9A–B). For a given interval $[i, j]$, we reconstruct the features using only the components in that band of the quantized basis:

$$\tilde{\mathbf{x}}_{[i,j]} = \boldsymbol{\mu}_Q + \sum_{k=i}^{j} \left((\mathbf{X}_Q - \boldsymbol{\mu}_Q) \cdot \mathbf{v}_{Q,k}\right) \mathbf{v}_{Q,k}^T \tag{10}$$

These partial reconstructions $\tilde{\mathbf{x}}$ are then fed into the model's original classification head (zero-shot text projection) to compute Top-1 accuracy. We report the relative accuracy change:

$$\Delta\text{Acc}_{[i,j]} = \frac{\text{Acc}_Q([i,j]) - \text{Acc}_{FP32}([i,j])}{\text{Acc}_{FP32}([i,j])} \tag{11}$$

Normalized Spectral Energy Change. We analyze the redistribution of information density (Figure 9C). We compute the Cumulative Distribution Function (CDF) of the explained variance for the quantized model:

$$F_Q(k) = \frac{\sum_{i=1}^{k} (\sigma_Q^{(i)})^2}{\sum_{j=1}^{D} (\sigma_Q^{(j)})^2} \tag{12}$$

The plot displays the differential $F_Q(k) - F_{FP32}(k)$. A positive value indicates that a larger proportion of the model's total energy is concentrated in the earlier ranks than in the FP32 baseline, confirming the model's reliance on the lower ranks. Note that the SVD ranks are recomputed for each model.

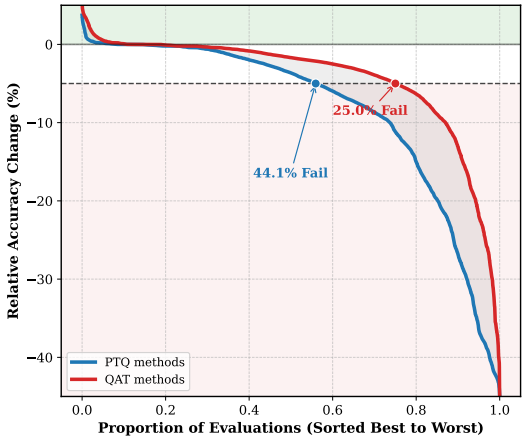

Figure 11. PTQ vs QAT overall success (60k runs).

Post-Training Quantization vs. Quantization-Aware Training in VLMs. Beyond average performance, reliability is defined by worst-case behavior. The Cumulative Distribution Function (CDF) in Figure 11 reveals that PTQ fails the 5% accuracy tolerance threshold in 44.1% of evaluations. Figure 12 offers a per-model landscape. QAT significantly acts as a stabilizer, nearly halving this failure rate to 25.0%. As detailed in Figure 14 (see Section E for a full breakdown), this reliability is maximized by the rotation-based method (Ashkboos et al., 2024), which essentially drastically reduces the quantization step by geometrically redistributing the huge magnitude of the outliers (Darcet et al., 2024).

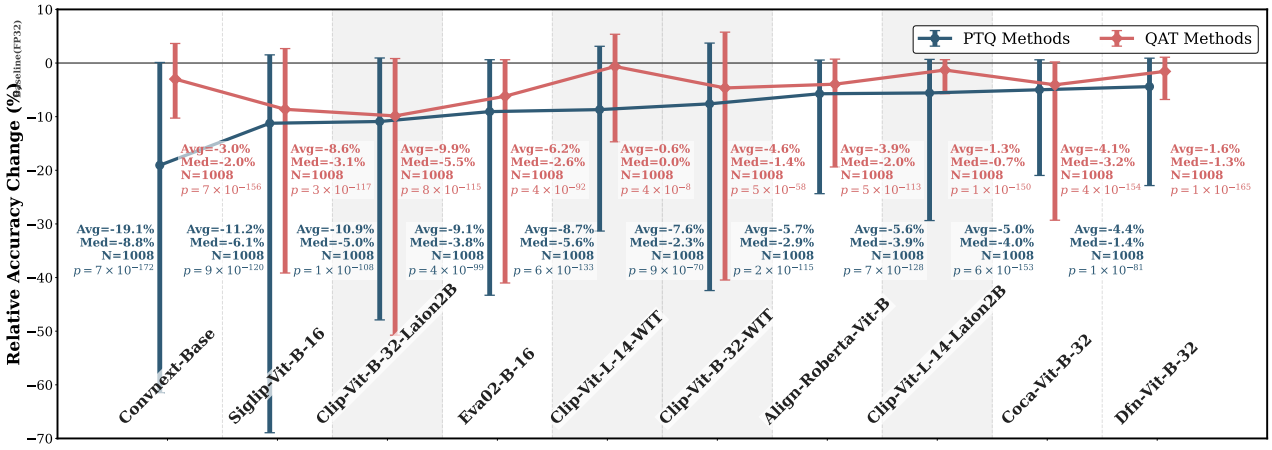

Figure 12. Average impact of post-training and quantization-aware training methods on zero-shot accuracy on several VLMs. Refer to the Appendix Figure 14 for the quantization-method success rate.

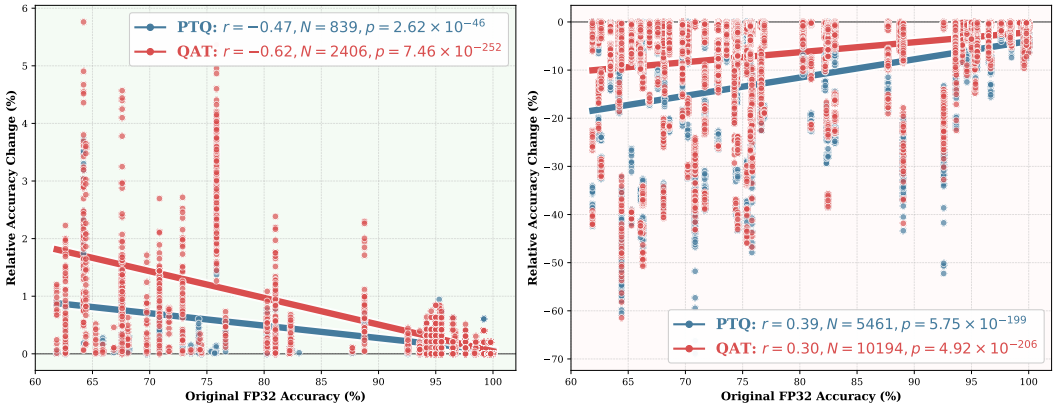

Figure 13. We plot quantization's impact on accuracy relatively to baseline accuracy and separate accuracy improvement (Left) from accuracy degradation (Right), and we notice a strong and significant correlation between the complexity of the task (low zero-shot accuracy) and quantization impact on it.

D. Task complexity imposes redundancy

The statistical landscape in Figure 13 reveals that quantization impact is not uniform but scales with task difficulty. In "easy" regimes (FP32 Accuracy $> 90\%$), decision boundaries are robust enough to absorb precision loss with minimal variance. However, as task complexity increases, the model enters a volatile state where quantization either clarifies or destroys the underlying representation.

The Regularization Gain: In low-accuracy regimes, we observe the strongest relative improvements ($r = -0.62$, Left). By reducing precision, we force the model to discard brittle, high-precision details in favor of more stable features that generalize better on difficult benchmarks.

The Fragility Collapse: Conversely, these same difficult tasks suffer the steepest degradations when the model's structural margin (Figure 2) is exhausted. For architectures with low native redundancy like `SigLIP`, the noise of quantization becomes a destructive force that wipes out the fragile signal required to solve complex cases, confirming that quantization resilience is a function of both task difficulty and available model capacity.

E. Granular Analysis of Failure Modes and Catastrophic Forgetting

In this section, we provide a detailed breakdown of the failure rates summarized in Figure 14. We define a "failure" as a relative zero-shot accuracy drop exceeding 5% compared to the FP32 baseline on unseen target datasets. This distinction is critical: because our QAT methods utilize small "proxy" datasets (e.g., CC3M) for calibration, performance on the target benchmarks (e.g., ImageNet) serves as a direct proxy for the model's resistance to **catastrophic forgetting**. **Geometric Alignment Mitigates Outliers without Training Risks.** The dominant trend in our evaluation is the superiority of rotation-based preprocessing. In the PTQ

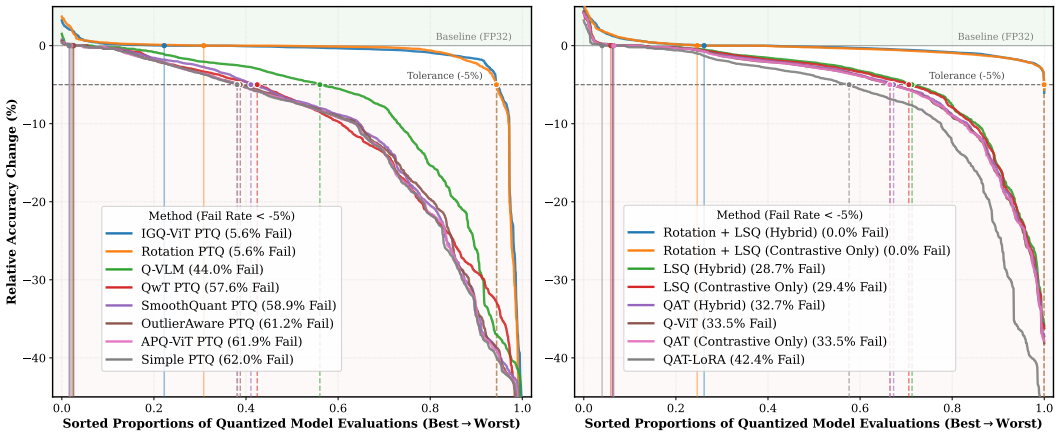

Figure 14. Relative accuracy drops using diffrent PTQ and QAT methods, we consider that a quantization has failed if the relative accuracy drop between the baseline model and quantized one on a test set is higher than 5%.

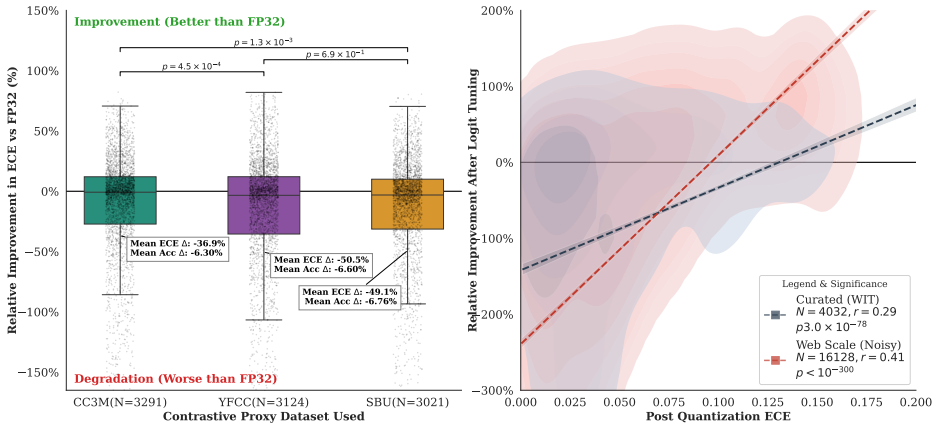

Figure 15. **Proxy Data Impact: For Calibration, Data Quality Matters, Unlike Accuracy** (Left) While accuracy is robust to proxy noise, calibration is highly sensitive; the curated CC3M dataset mitigates ECE loss significantly better than noisy alternatives. (Right) Severe post-quantization misalignment yields proportionally greater gains from logit-scale tuning, confirming that reliability can be repaired via logit-scale tuning.

regime (Figure 14, Left), `Rotation PTQ` achieves a failure rate of only 5.6%, vastly outperforming `Simple PTQ` (62.0% failure). By applying random orthogonal matrices to weight and activation spaces, these methods redistribute outlier features (”massive activations”) across multiple dimensions. Crucially, because this operation is strictly geometric and reversible, it renders the model ”quantization-friendly” without altering its semantic alignment. This preserves the model’s pure zero-shot capabilities, as no gradient updates are applied that could bias the model toward the calibration data. **The LoRA: Parameter Efficiency vs. Spectral Restoration.** A counter-intuitive finding is the high failure rate of `QAT-LoRA` (42.4%), which performs worse than standard full-parameter QAT. While Low-Rank Adaptation is typically prized for preventing catastrophic forgetting, in the context of quantization, it appears structurally insufficient. Quantization noise is high-frequency and distributed across the entire spectral spectrum of the weight matrices. LoRA, by definition, restricts updates to a low-rank subspace. Our results suggest that LoRA lacks the degrees of freedom required to compensate for the pervasive noise of low-bit quantization. Consequently, the adapter overfits to the proxy calibration set in a futile attempt to recover performance, leading to significant degradation on zero-shot tasks (catastrophic forgetting).

Hybrid Robustness: The Synergy of Rotation and LSQ. The most robust configuration observed is `Rotation + LSQ (Hybrid)`, which achieves a perfect 0.0% failure rate. This method combines the best of both worlds:

1. **Rotation** pre-conditions the weight space, flattening outliers and reducing the ”quantization error budget” before training begins.

2. **LSQ** (Learned Step Size Quantization) performs minimal, targeted updates to the scaling factors rather than

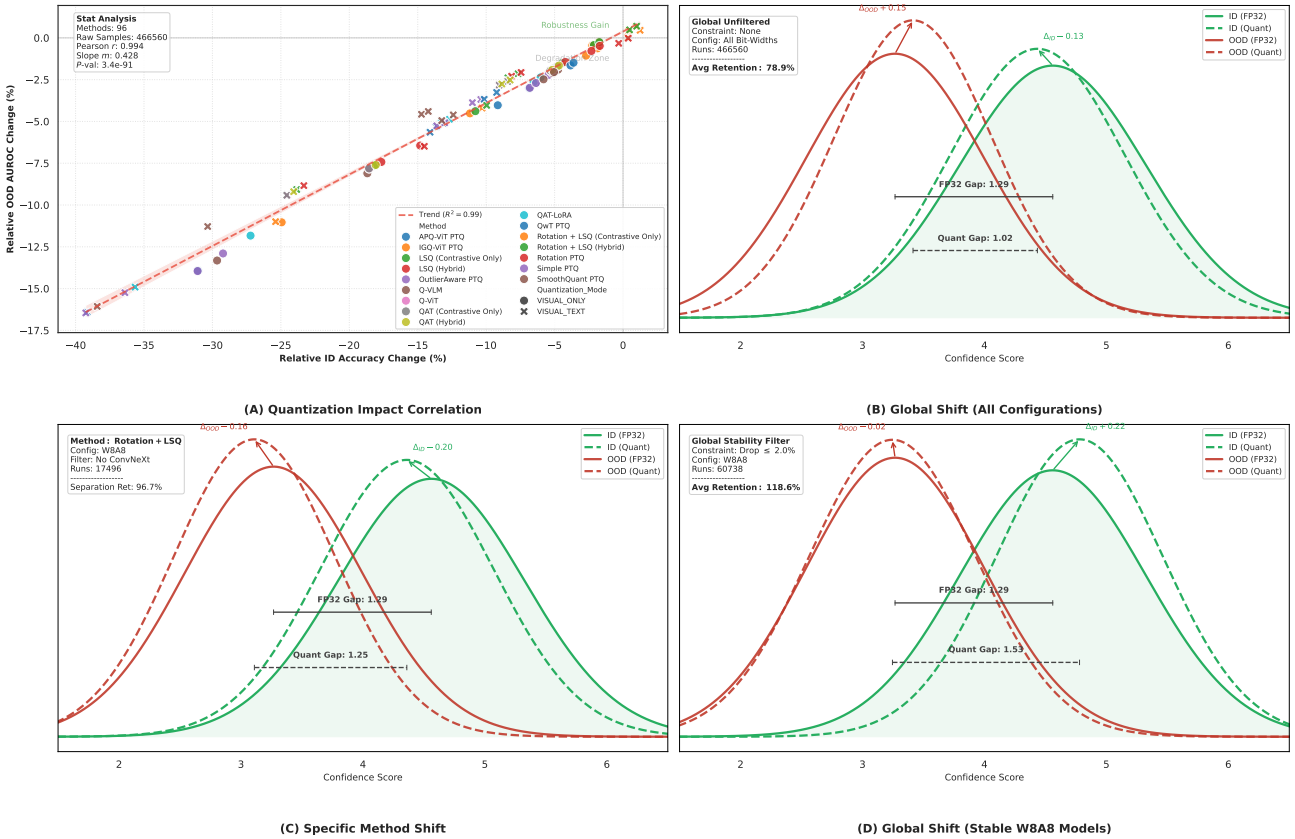

Figure 16. **Mechanics of Quantization-Induced Separability.** (A) Global correlation analysis ($N = 466k$) reveals a shallow slope ($m = 0.428$), indicating OOD detection is more robust than accuracy. (B-D) Kernel Density Estimates of confidence scores. While aggressive quantization causes distribution collapse (B), for (C), the preservative power of rotation push both the ID and OOD distribution as quantization noise is added, maintaining separability as accuracy/confidence degrades, while stable W8A8 models (D) exhibit "Manifold Contraction," where ID samples anchor to quantized centroids while OOD samples remain low-confidence, effectively increasing the separability gap (Δ_{Gap}) from 1.29 to 1.53.

the weights themselves.

By solving the geometric problem analytically (via rotation) and the precision problem parametrically (via LSQ), this hybrid approach minimizes the magnitude of gradient updates required. This drastically reduces the risk of the model drifting away from its pre-trained manifold, thereby preventing catastrophic forgetting and maintaining robust pure zero-shot generalization.

F. Mechanisms of OOD Separability under Quantization

To explain why OOD detection remains robust despite precision loss, we analyze the distributional dynamics of confidence scores for In-Distribution (ID) and Out-of-Distribution (OOD) samples.

Quantization-Induced Manifold Contraction. The kernel density estimates in Figure 16 visualize the impact of quantization on score distributions. In the FP32 baseline, the separability gap between the mean ID and OOD confidence is 1.29. The effect of quantization varies by model stability:

1. **Destructive Regime (Panel B):** When quantization is aggressive or unoptimized, the representation degrades significantly. Both ID and OOD distributions shift toward lower confidence, and their variances overlap, reducing separability.

2. **Stable Regime (Panel D):** In successful W8A8 quantization (defined by $\leq 2\%$ accuracy drop), we observe a *widening* of the separability gap to 1.53.

F.1. Quantization's Impact on ConvNeXt: A Spectral Perspective

While Vision Transformers (ViTs) exhibit stability under quantization, the `ConvNeXt` architecture demonstrates a

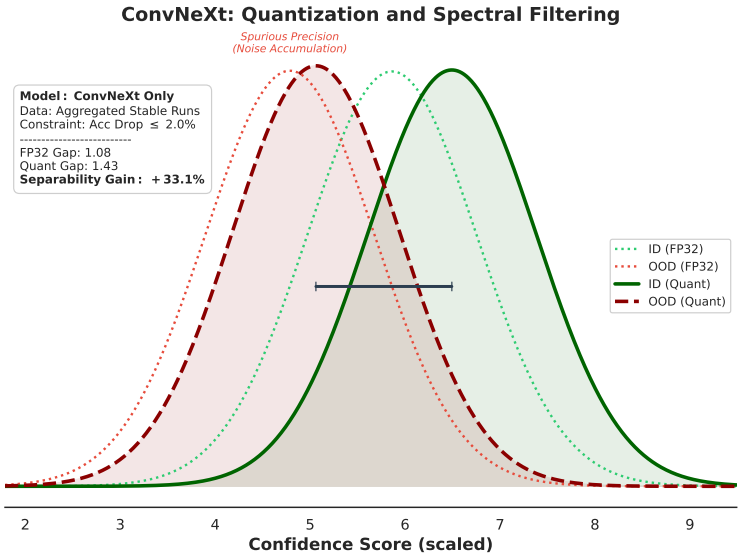

Figure 17. **Quantization as a Spectral Filter for ConvNeXt.** Visualization of the aggregated In-Distribution (ID) and Out-of-Distribution (OOD) score distributions for `ConvNeXt-Base`. Unlike standard compression degradation, quantization increases the ID-OOD separability gap by **33.1%** relative to the FP32 baseline (Ghost lines). The *Spectral Filtering* mechanism (Section F.1) effectively dampens the high-frequency "spurious precision" (Red arrow) that drives OOD overconfidence in FP32, while the robust low-rank core anchors the ID confidence (Green arrow), verifying the hypothesis that discretization acts as a beneficial denoiser for CNNs.

distinct positive response, showing a 33.1% relative improvement in OOD separability (Figure 17). Using the spectral framework established in Section 5, we can now attribute this to the interaction between Passive Spectral Filtering and the specific inductive biases of Convolutional Neural Networks (CNNs).

The Spectral Bias of CNNs. Unlike the global attention mechanism of ViTs, which naturally promotes low-rank, semantically smoothed representations (shape bias), CNNs exhibit a strong bias toward high-frequency, local textural features. In FP32 precision, `ConvNeXt` utilizes this "spectral tail" (High Ranks >64) to maximize training accuracy. However, for OOD detection, this tail is toxic: it allows the model to assign high confidence to OOD samples based on superficial textural similarities (e.g., a "dog" texture on a non-dog object), resulting in the congested logit space observed in the FP32 baseline (Gap: 1.08).

Quantization as a Low-Pass Filter. As demonstrated in Figure 8c, quantization causes a collapse in the Signal-to-Quantization-Noise Ratio (SQNR) for high-rank, low-variance components. For ConvNeXt, this acts as a beneficial **spectral filter**:

1. **Tail Truncation:** The quantization grid effectively acts as a low-pass filter, mathematically preventing the model from encoding the fine-grained, high-frequency textural features that drive OOD overconfidence.
2. **Forced Shape Bias:** With the "textural shortcut" removed, the model is compelled to rely on the **Active Subspace Concentration** described in Figure 9A. The decision boundary shifts to rely almost exclusively on the robust, low-rank semantic core (Ranks 0-8), which encodes global shape and semantic structure.

Resulting Topography. As visualized in Figure 17, this spectral purification increases the separability gap to 1.43. By stripping away the high-rank noise that CNNs are prone to overfitting, quantization ironically acts as a regularizer that aligns the ConvNeXt latent space more closely with the robust, low-rank topology typical of ViTs.

F.2. Quantization Sample Size Selection

To determine the optimal calibration dataset size for our quantization pipeline, we performed a scaling study ranging from 1 to 10,000 unique samples across three proxy datasets (CC3M, YFCC, SBU). As illustrated in Figure 18, our analysis identifies $N = 1000$ as the critical sample budget.

In the low-data regime ($N < 1000$), standard methods like LSQ suffer significant degradation (up to -4% accuracy drop on ImageNet), as they fail to estimate stable step-size parameters from sparse statistics. Conversely, extending the calibration set beyond this point ($N > 1000$) yields diminishing returns and introduces the risk of catastrophic forgetting. Because we employ Quantization-Aware Training (QAT) on proxy data that is distributionally distinct from

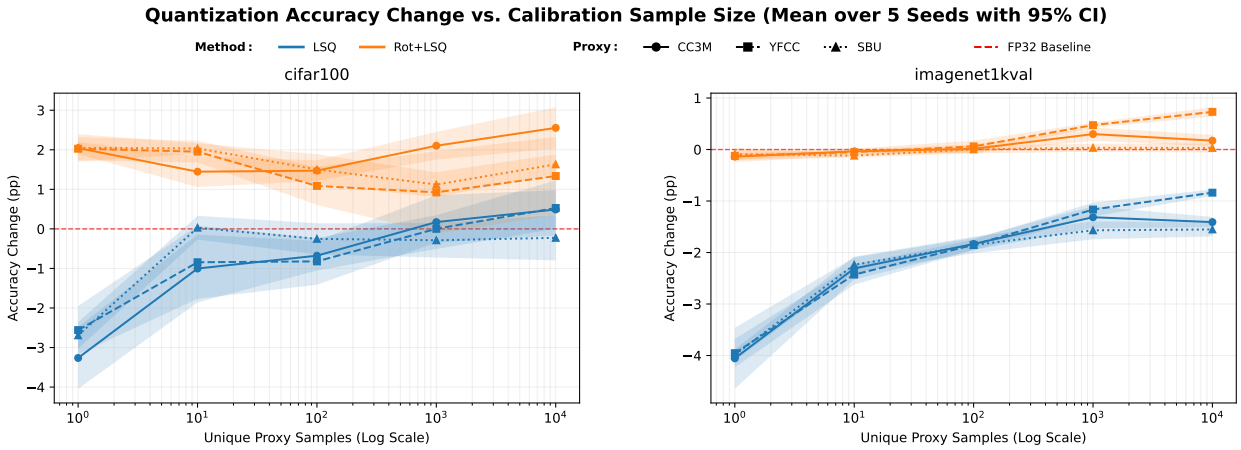

Figure 18. Sensitivity Analysis of Calibration Sample Size: We evaluate the impact of the number of unique proxy samples ($N \in [1, 10^4]$) on Zero-Shot Accuracy relative to the FP32 baseline. The shaded regions represent the 95% confidence interval over 5 random seeds. While *Rotation+LSQ* (Orange) demonstrates high robustness at low sample regimes, standard *LSQ* (Blue), while surprisingly effective with the statistics of a single image, continues to converge when fed with more unique samples. We select $N = 1000$ as the optimal operating point to balance quantization effectiveness against the risks of overfitting to the proxy distribution and catastrophic forgetting (Kumar et al., 2022).

the target benchmarks, excessive training on large proxy subsets causes the model parameters to overfit the proxy domain, drifting away from the generalizable features of the pre-trained FP32 manifold. Therefore, $N = 1000$ represents the requisite trade-off: it provides sufficient diversity for the convergence of data-hungry methods like LSQ while preventing the distributional drift associated with large-scale proxy fine-tuning.

F.3. Simulated Quantization

A crucial aspect of our methodology is the use of simulated (or 'fake') quantization. This approach is strictly necessary for several converging reasons. First, hardware limitations: Current GPUs lack native arithmetic support for the non-standard bit-widths we evaluate (e.g., INT6, INT4) and often lack support for fully quantized activations in Transformer architectures, for our comparisons to be valid, they require to apply the same quantizations to all models and architectures, which would be impossible without simulations, as not all architecture-specific layers have quantized kernels. Second, isolation: Simulation allows us to isolate the theoretical impact of precision loss (spectral filtering) from hardware-specific implementation quirks (e.g., kernel overflow or accumulation limits, layers fusion, etc..), ensuring our findings regarding reliability are fundamental to the compression method rather than the hardware backend. Finally, while libraries like bitsandbytes offer efficient kernels for weight-only quantization, they do not support the comprehensive weight-and-activation quantization required to model the full spectral regularization effects observed in this study.

Verification of Simulation Correctness. To ensure our simulation was valid, we implemented a verification utility. For any quantized model, this utility counts the number of unique values in both the weight and activation tensors (via forward hooks). For a successful INT8 simulation, this count must be $\leq 2^8 = 256$, which we confirmed for our implementations.

Table 2. Performance and memory benchmark of OpenAI CLIP models using non-simulated, kernel-based quantization (bitsandbytes). Inference time is the average latency per batch on an RTX 3090 GPU. Note the significant slowdown for 8-bit and 4-bit inference, which motivates our use of simulation to evaluate accuracy/reliability trade-offs, as current kernels are not optimized for inference speed. This table highlights why we must use simulated quantization and not confuse general quantization frameworks like QLoRA, bitstandbytes that quantize to train, with our aim of lightweight deployment.

Model	Prec.	Mem. (MB)	Time (ms)	Speed (×)	Gain (%)
B/32	FP32	358.3	19.6	1.00	0.0
	FP16	297.8	17.7	1.11	16.9
	INT8	179.8	81.1	0.24	49.8
	NF4	123.8	47.7	0.41	65.4
B/16	FP32	579.5	12.3	1.00	0.0
	FP16	294.6	14.2	0.87	49.2
	INT8	177.7	90.9	0.14	69.3
	NF4	119.7	49.8	0.25	79.3

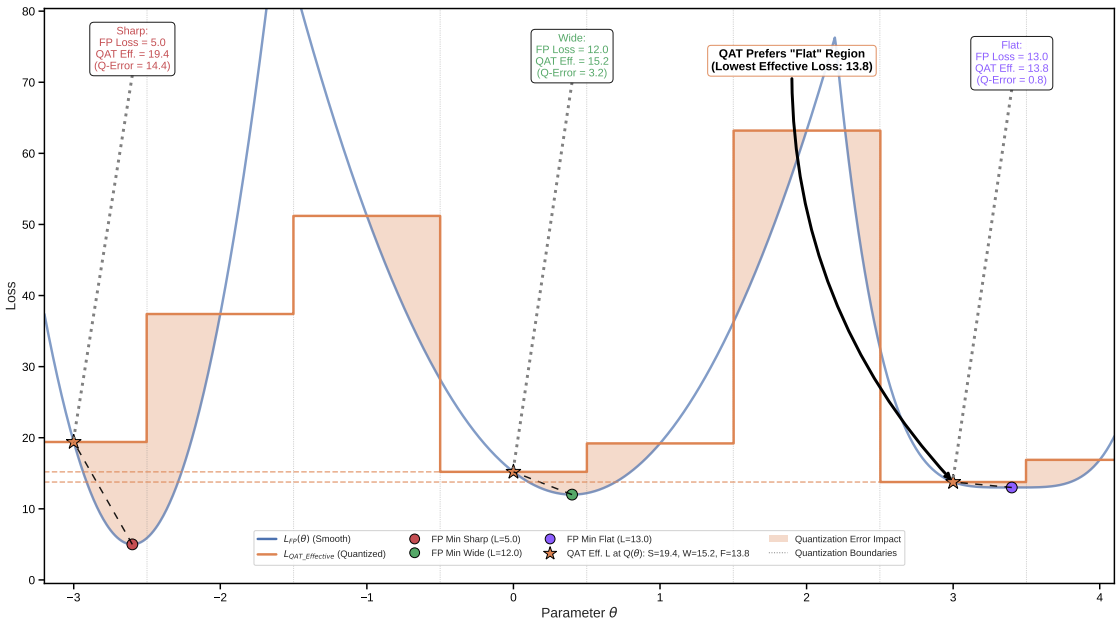

Figure 19. A conceptual illustration of how QAT forces the optimizer to abandon a sharp minimum in favor of a flatter, more robust solution that has less loss under quantization.

F.4. Motivation: Why Not Use Existing Kernels like QLoRA, bitsandbytes for Inference?

While methods like QLoRA enable the *training* of models with 4-bit weights, they are primarily designed to reduce memory usage during the training phase, not to accelerate inference. As shown in our direct performance benchmark in Table 2, using these kernels for inference can lead to a significant *slowdown* compared to the FP32 baseline. The overhead of de-quantizing the weights on-the-fly for each computation outweighs the benefits of reduced memory bandwidth. True inference acceleration requires dedicated hardware and software support for low-precision matrix multiplications for multi-head attention, which is what our simulation-based study aims to evaluate in terms of potential accuracy and reliability before such support becomes widespread.

F.5. The core functioning of Fake Quantization

Fake quantization is a simulation technique that models the error introduced by quantization and de-quantization within a standard full-precision (FP32/FP16) training and inference loop. The process for a tensor x is as follows:

1. **Quantize:** The full-precision input tensor x is scaled, shifted, and rounded to the nearest integer value within the target bit-width's range (e.g., [-128, 127] for INT8).

$$x_{quant} = \text{round}\left(\frac{x}{\text{scale}} + \text{zero\_point}\right)$$

2. **Clamp:** The integer values are clamped to the representable range of the target bit-width.

3. **De-quantize:** The clamped integer tensor is immediately converted back to a full-precision floating-point tensor.

$$x_{dequant} = (x_{quant} - \text{zero\_point}) \times \text{scale}$$

The resulting tensor, $x_{dequant}$, has the same data type as the input but contains the precision loss that *would have occurred* in a true low-bit system. For the backward pass, a **Straight-Through Estimator (STE)** (Courbariaux et al., 2015) bypasses the non-differentiable 'round' function, enabling the model to learn weights that are robust to the simulated quantization noise.

G. On QAT's Preference for Flat Minima

A significant body of work establishes that neural networks converging to "flat" minima in the loss landscape generalize better (Hochreiter & Schmidhuber, 1997). More recently, (Saqib et al., 2025) has shown that weight-only quantization-aware training can improve domain generalization by leveraging this phenomenon. As illustrated in Figure 19, QAT inherently penalizes sharp regions because the weight perturbation from quantization ($Q(w) - w$) causes a large increase in loss. This forces the optimizer to seek out flatter, more robust regions, which can sometimes lead to better generalization and accuracy than the original FP32 model. To be comprehensive, we illustrate the discrepancy between the theoretical full-precision optimization landscape, denoted

by the continuous curve $\mathcal{L}_{\text{FP}}(\theta)$, and the effective loss realized under discrete quantization constraints, represented by the stepped function $\mathcal{L}_{\text{QAT}}(\theta)$. The landscape features three distinct minima characterized by varying local curvature (Hessian magnitude): a "Sharp" global minimum, a "Wide" intermediate basin, and a "Flat" local minimum. While the Sharp region ($\theta \approx -2.8$) achieves the lowest theoretical loss ($\mathcal{L}_{\text{FP}} = 5.0$), its high curvature renders it hypersensitive to the parameter perturbations δ introduced by the quantization grid, resulting in a substantial discretization error (Q-Error ≈ 14.4). In contrast, the "Flat" region ($\theta \approx 3.5$), despite a higher baseline loss ($\mathcal{L}_{\text{FP}} = 13.0$), exhibits superior robustness to quantization noise. Due to the low gradient magnitude $\nabla_\theta \mathcal{L}$ and minimal curvature in this region, the quantization grid aligns more favorably with the loss surface, yielding the lowest *effective* quantized loss ($\mathcal{L}_{\text{QAT}} = 13.8$). This demonstrates that quantization-aware training implicitly favors flat minima, as they provide the necessary tolerance for the discrete mapping of parameters where $\mathcal{L}(\lfloor\theta\rceil) \approx \mathcal{L}(\theta)$.

CLIP Quantization and Fairness Previous work on CLIP's quantization (Slyman et al., 2024) has revealed small but generally non-significant increases in unfair biases. Note that they infer using non-deployable quantization frameworks, such as "HuggingFace 4-bit Quantization," which uses QLoRA (quantization for training larger models with less memory) but does not increase inference speed (as tested 2) and does not quantize activations. "Pytorch 8-bit Quantization," Which only supports CPU PTQ methods. And "LLM.int8()", which does not quantize activations. We highlight the need for engineering low-precision kernels for CLIP that truly aim for lightweight deployment.

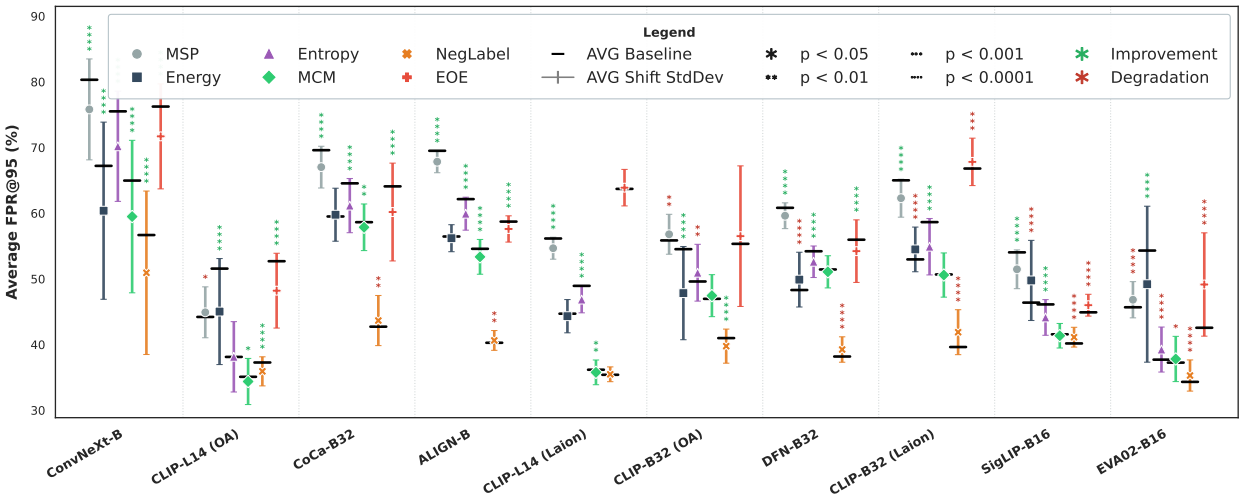

Figure 20. Impact of successful W8A8 quantization on OOD Detection (FPR@95). Average FPR@95 across quantization methods (lower is better). QAT methods (center, right) maintain OOD performance for the LAION model, despite this model suffering from significant accuracy and calibration degradation. VLM-specific OOD methods consistently outperform classic methods. Vertical lines represent the maximum improvement and degradation relative to the full precision baseline of that experiment.

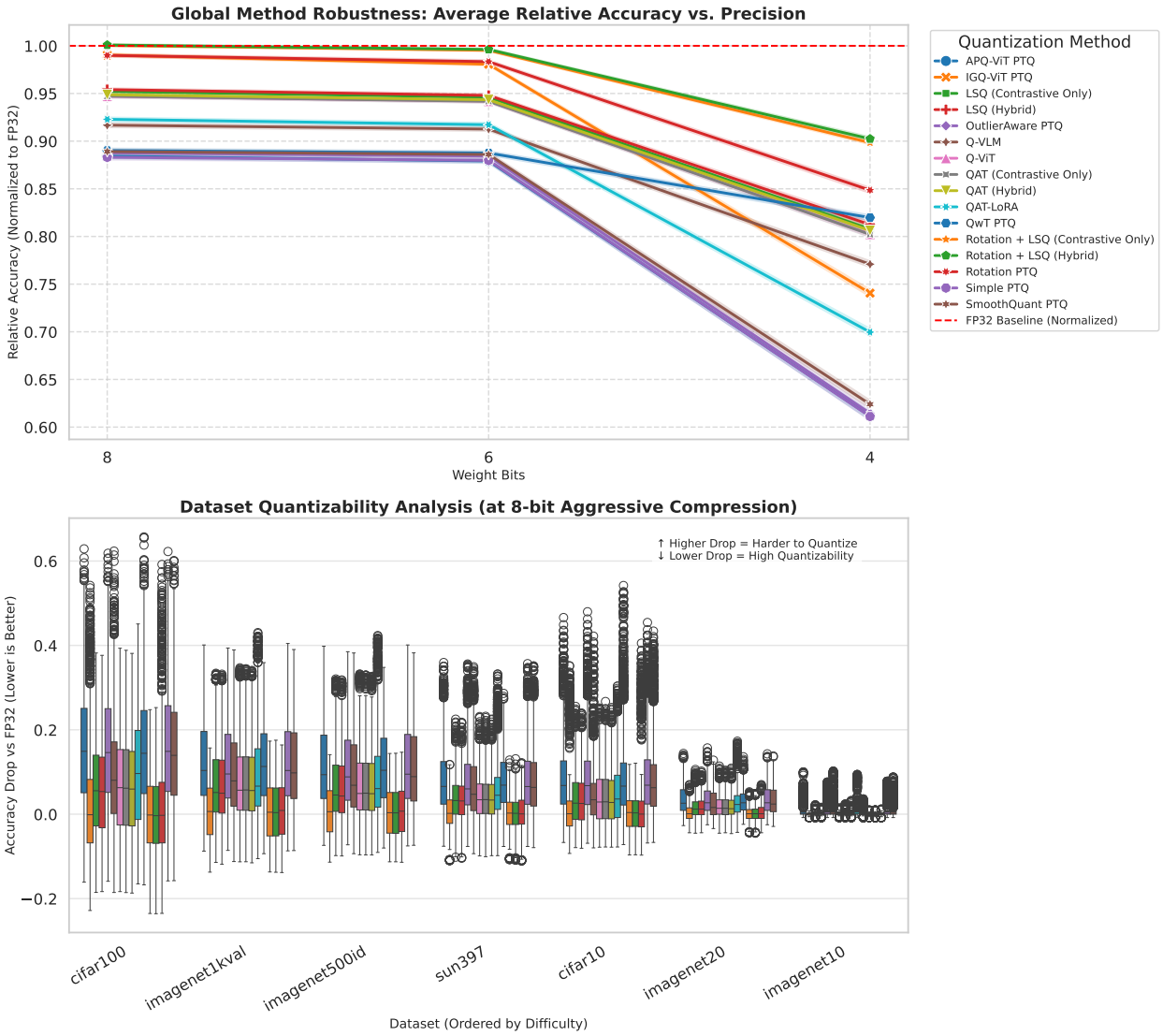

Figure 21. (Top) Global accuracy retention per quantization method relative to increasingly aggressive quantization. (Bottom) Per the downstream dataset, zero-shot evaluation after quantization (aggregation of 60k data points) to show that low-resolution datasets like CIFAR-100 benefit more because they are more coarse-grained and feature-friendly. (0.6=60%)

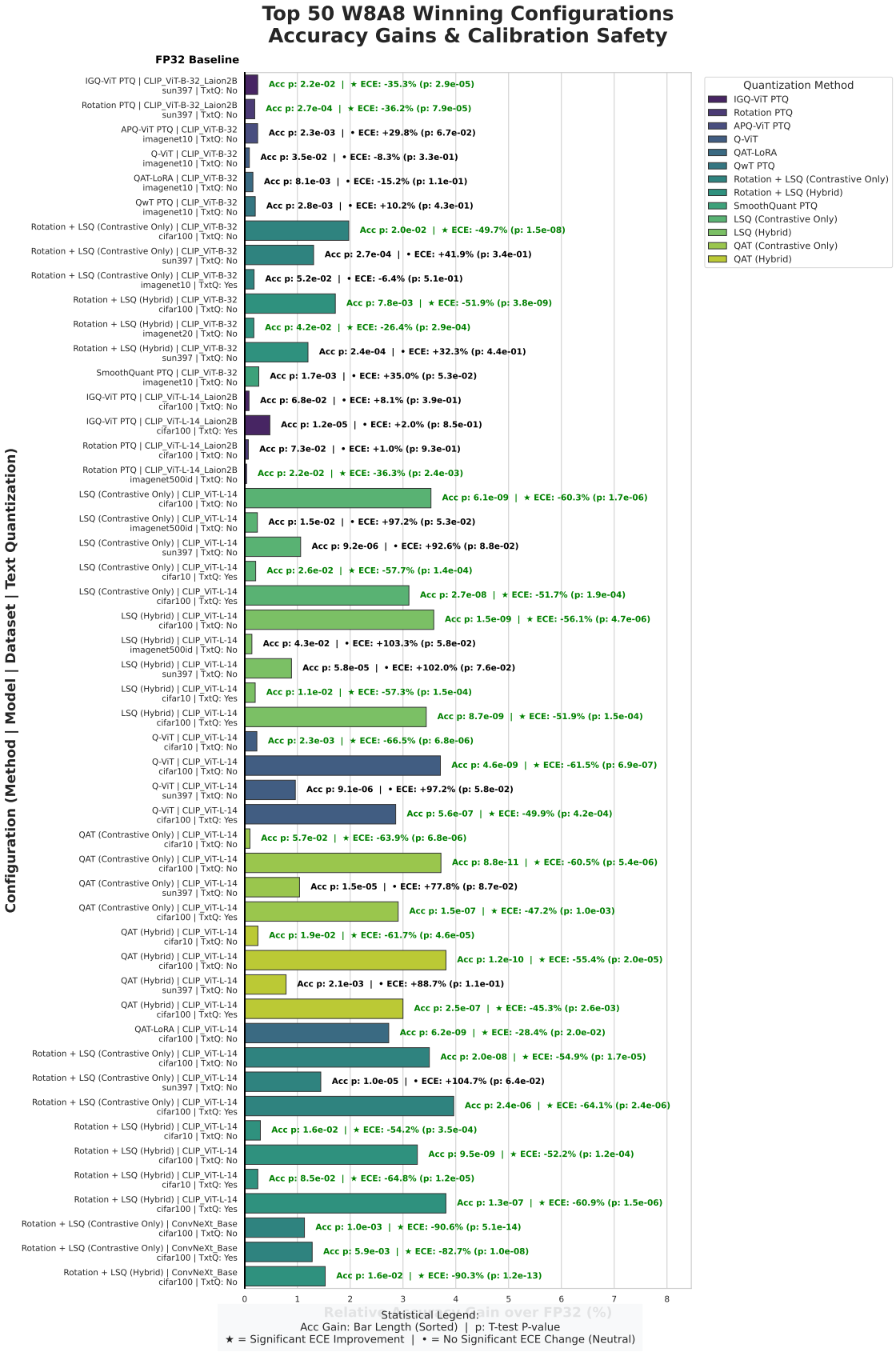

Figure 22. Accuracy-Winning configurations (showing only 50). Green text indicates a simultaneous improvement in calibration. While advanced PTQ methods (IGQ, ROT) often show only slight improvement, QAT methods improve accuracy less frequently but more drastically.

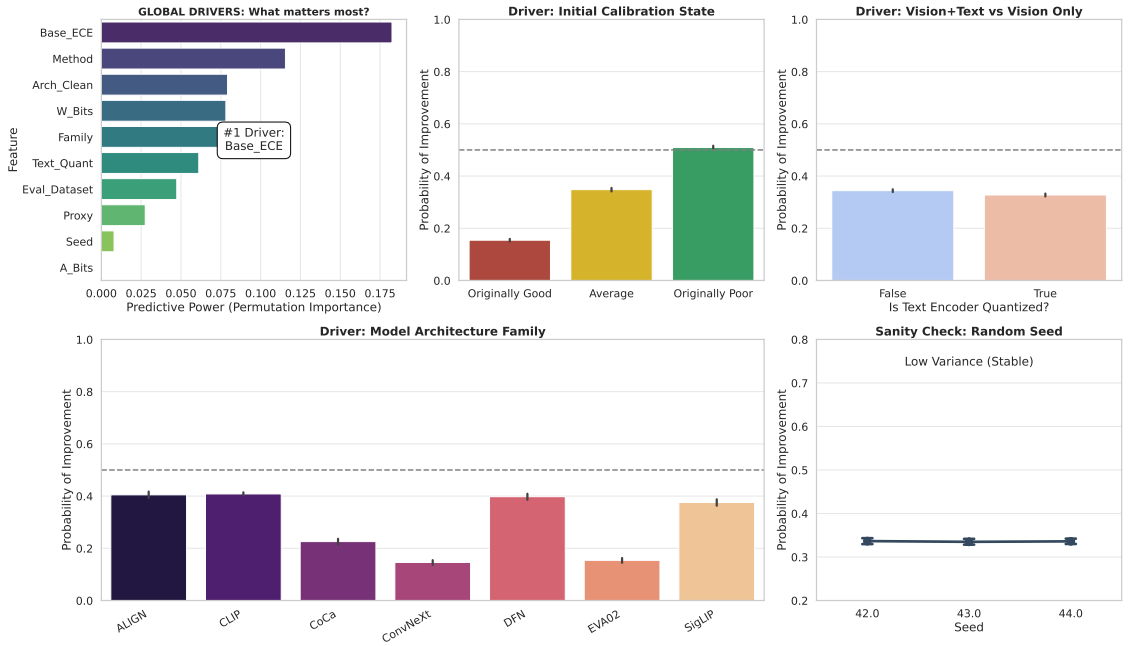

Figure 23. What parameters improve calibration? (global view)

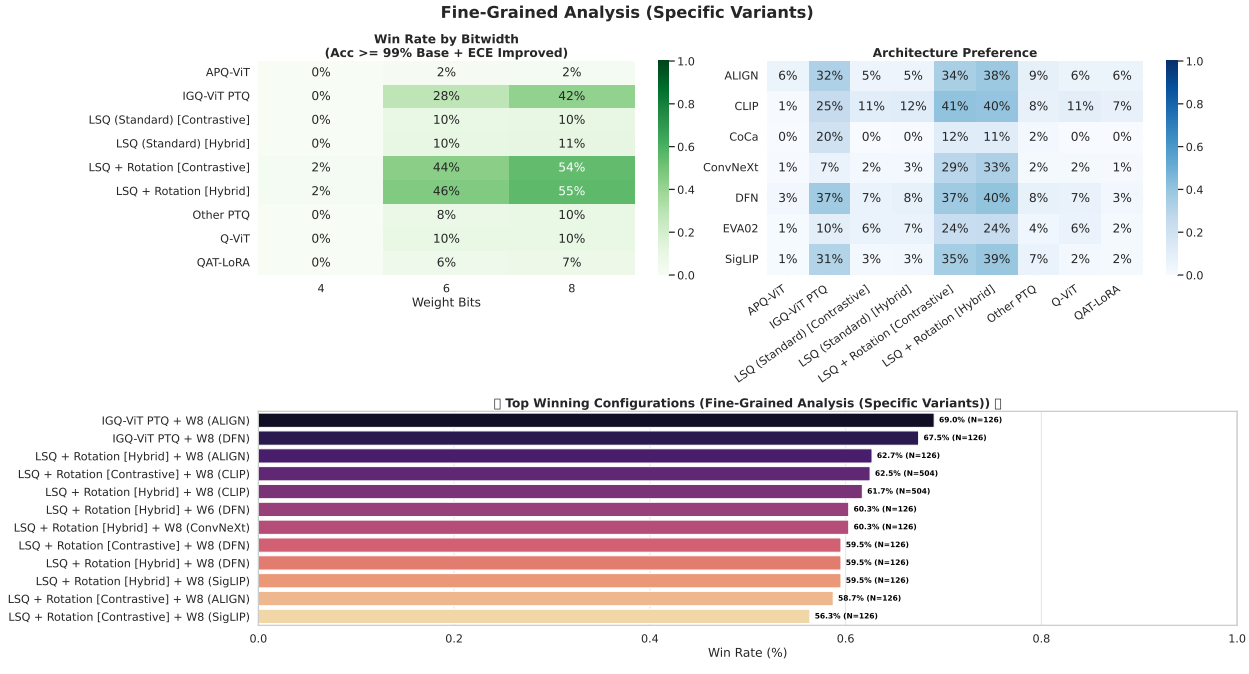

Figure 24. What parameters improve calibration? (Closer view)

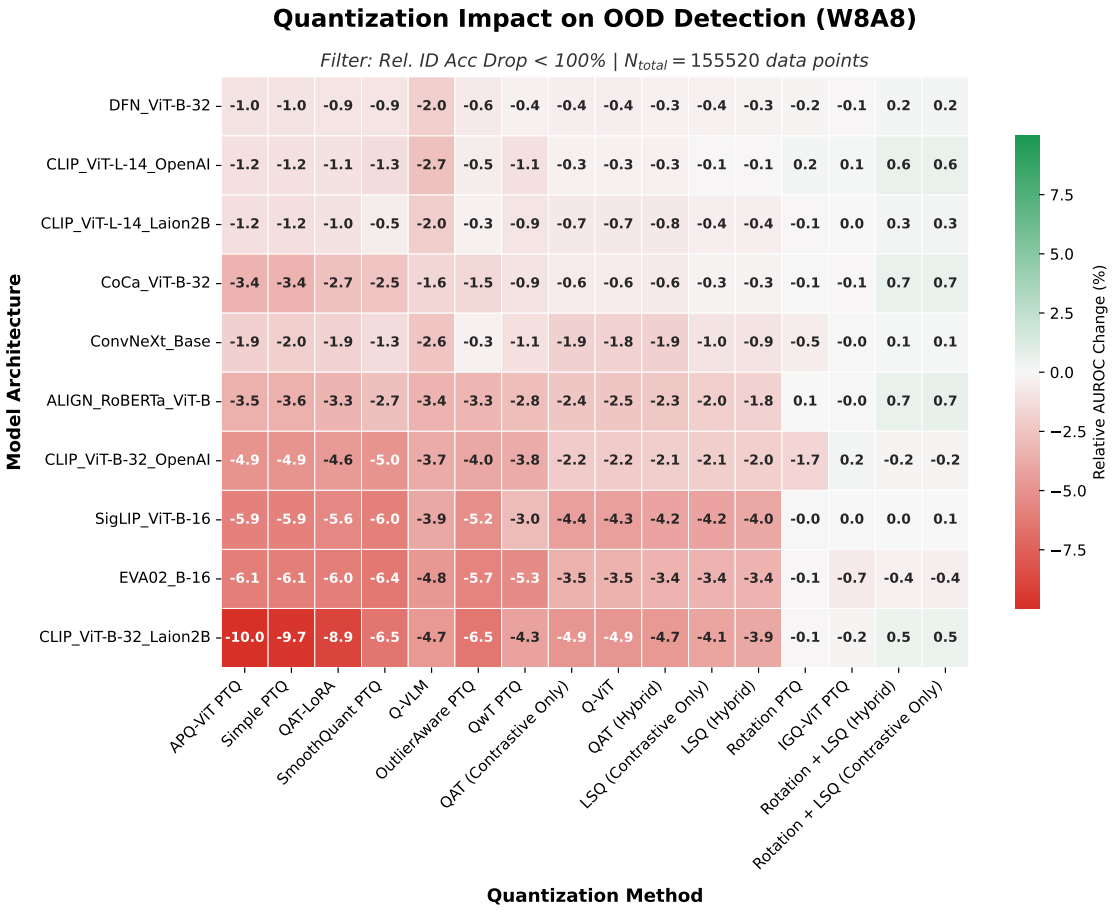

Figure 25. Global impact of all W8A8 quantization methods on models (OOD detection methods scores are aggregated).

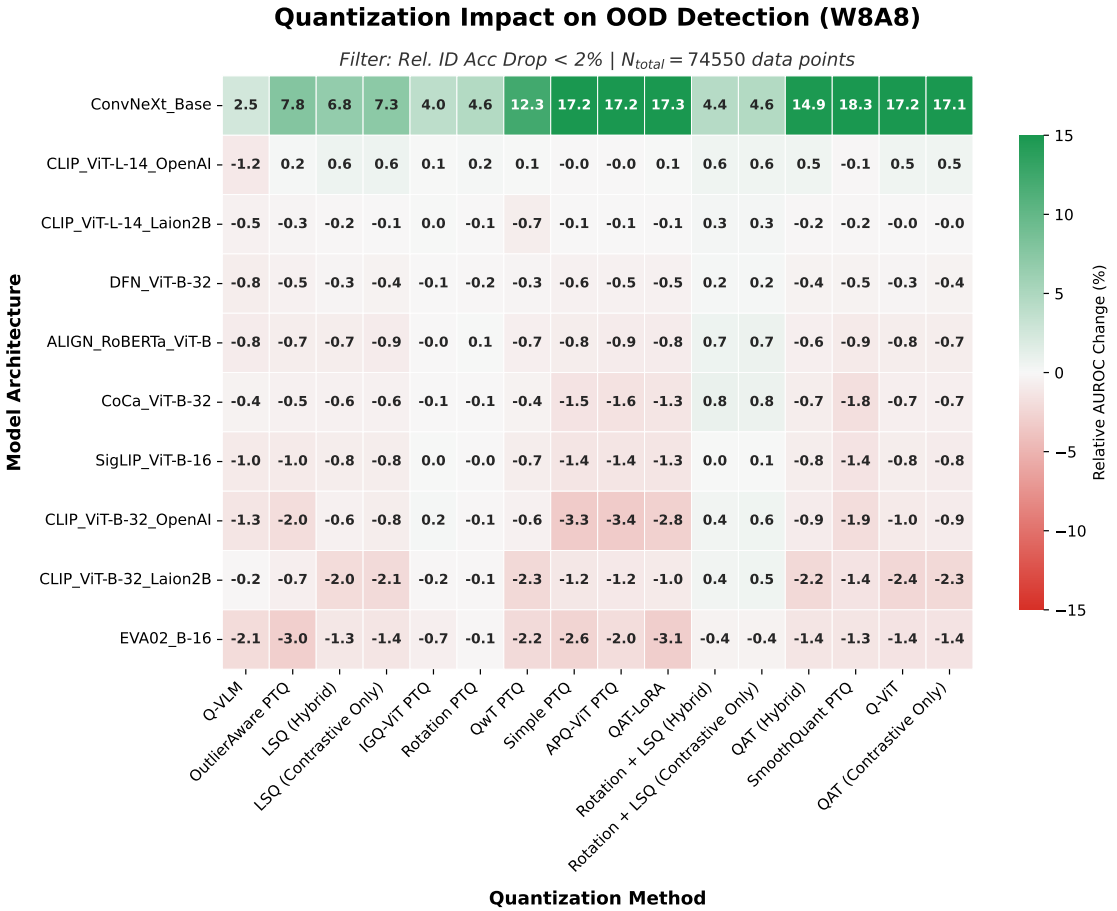

Figure 26. Global impact of successful W8A8 quantization (less than 2% relative degradation on the ID accuracy) on models (OOD detection methods scores are aggregated.

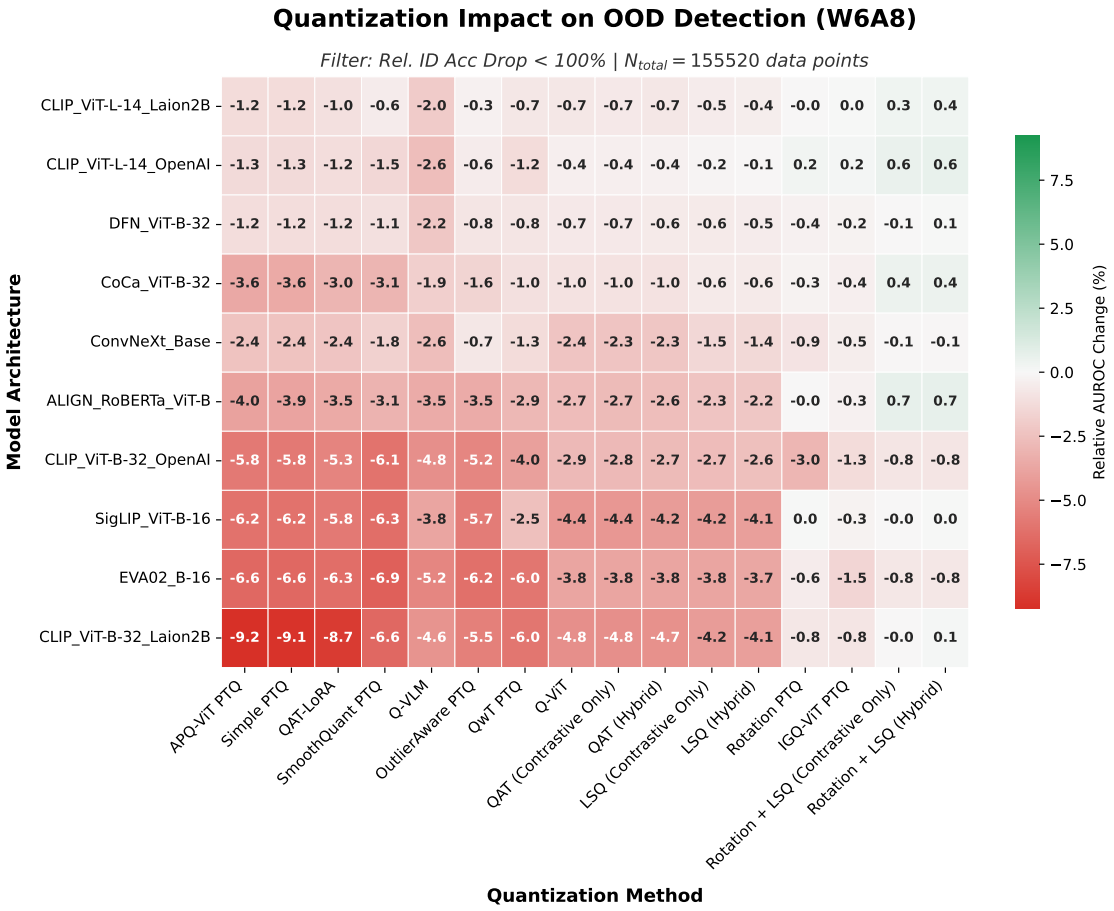

Figure 27. Global impact of all W6A8 quantization methods on models (OOD detection methods scores are aggregated).

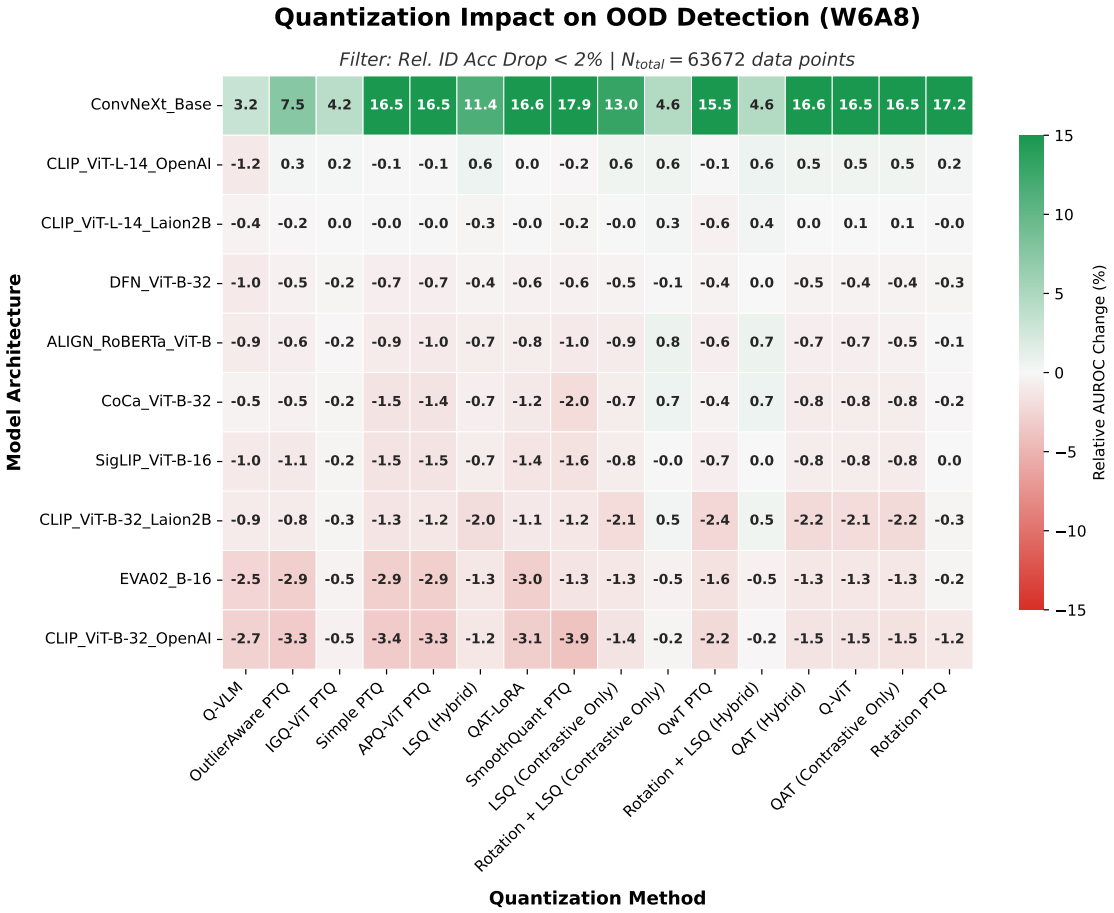

Figure 28. Global impact of successful W6A8 quantization (less than 2% relative degradation on the ID accuracy) on models (OOD detection methods scores are aggregated.

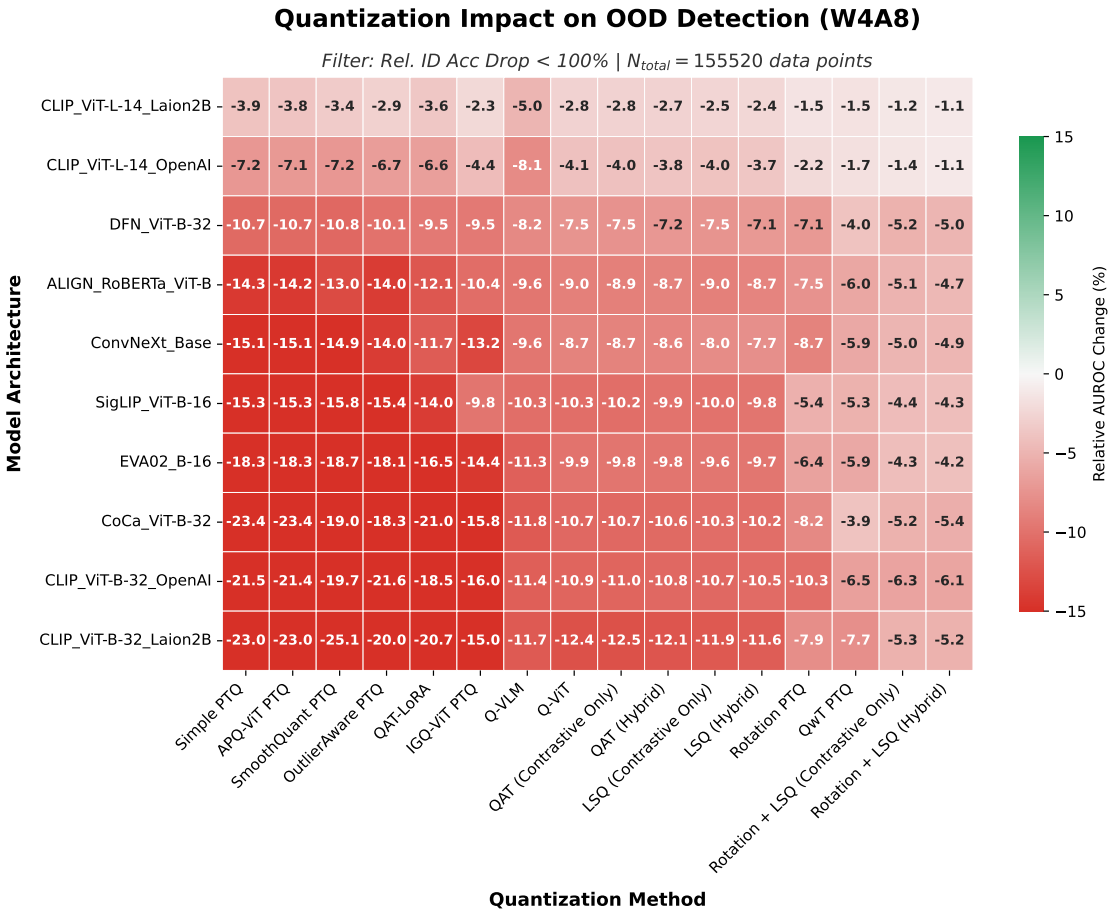

Figure 29. Global impact of all W4A8 quantization methods on models (OOD detection methods scores are aggregated).

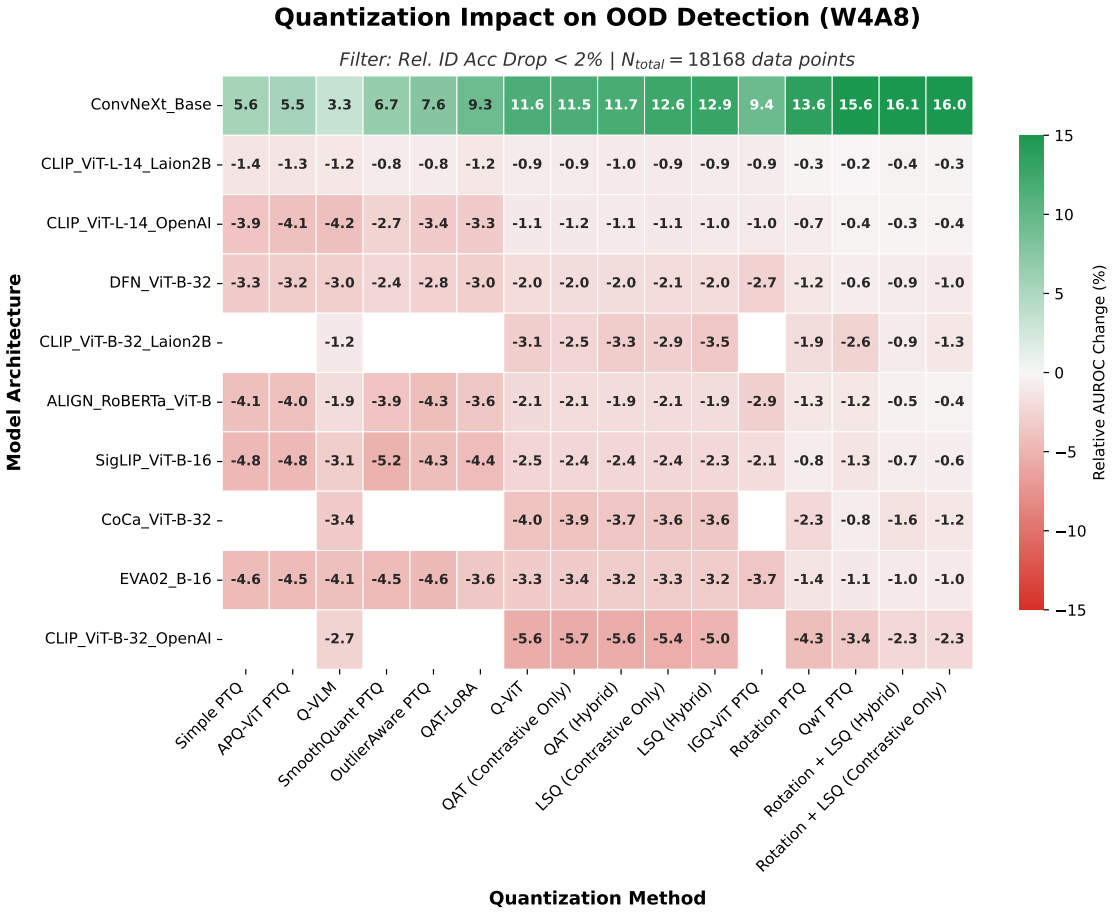

Figure 30. Global impact of successful W4A8 quantization (less than 2% relative degradation on the ID accuracy) on models (OOD detection methods scores are aggregated. (Empty squares mean that no sample exists)